

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

<http://go.warwick.ac.uk/wrap/874>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

# List-length and list-strength effects in recognition memory

Luciano Grüdtner Buratto

Submitted for the degree of Doctor of Philosophy

University of Warwick  
Department of Psychology  
July 2008

# Table of Contents

TABLE OF CONTENTS.....	I
LIST OF FIGURES .....	V
LIST OF TABLES .....	VII
LIST OF ABBREVIATIONS .....	VIII
DECLARATION.....	IX
ACKNOWLEDGMENTS .....	X
ABSTRACT .....	XI
CHAPTER 1. REVIEW AND OBJECTIVES .....	1
1.1. INTRODUCTION .....	1
1.2. EARLY GLOBAL MATCHING MEMORY MODELS .....	2
1.2.1. SAM ( <i>Gillund &amp; Shiffrin, 1984</i> ).....	5
1.2.2. MINERVA2 ( <i>Hintzman, 1988</i> ) .....	6
1.2.3. TODAM ( <i>Murdock, 1982</i> ).....	7
1.3. EMPIRICAL EVIDENCE: EARLY STUDIES .....	8
1.3.1. Evidence for list-length effect.....	8
1.3.2. Evidence against list-strength effect .....	10
1.3.3. Other challenges to global matching models.....	14
1.4. RECENT GLOBAL MATCHING MEMORY MODELS .....	15
1.4.1. SAM and the differentiation assumption.....	15
1.4.2. TODAM and the continuous memory assumption .....	16
1.4.3. REM ( <i>Shiffrin &amp; Steyvers, 1997</i> ) .....	17
1.5. EMPIRICAL EVIDENCE: RECENT STUDIES .....	20
1.5.1. Evidence against list-length effects.....	21
1.5.2. Evidence for list-strength effects.....	25
1.5.3. Recent challenges to global matching models: the role of recall ....	28
1.6. ALTERNATIVES: CONTEXT-NOISE AND DUAL-PROCESS MODELS .....	37
1.6.1. BCDMEM ( <i>Dennis &amp; Humphreys, 2001</i> ).....	37
1.6.2. SAC ( <i>Reder et al., 2000</i> ) .....	42

1.6.3. CLS (Norman & O'Reilly, 2003).....	48
1.7. AIMS OF THE THESIS .....	55
1.7.1. Empirical objectives.....	56
1.7.2. Theoretical objectives .....	59
<b>CHAPTER 2. GENERAL METHODOLOGY .....</b>	<b>62</b>
2.1. INTRODUCTION .....	62
2.2. SIGNAL DETECTION THEORY .....	63
2.2.1. Familiarity distribution.....	64
2.2.2. Receiver Operating Characteristic .....	68
2.2.3. Sensitivity measures ( $d'$ , $d_a$ , $A_z$ ).....	75
2.2.4. Bias measures ( $X_b$ , $c$ , $c_a$ ).....	78
2.3. DATA ANALYSIS .....	80
2.3.1. Raw measures .....	80
2.3.2. Derived measures.....	82
2.3.3. Power analysis .....	89
2.3.4. Regressions ( $zROC$ ).....	92
<b>CHAPTER 3. EXPERIMENTS 1-4 .....</b>	<b>94</b>
3.1. INTRODUCTION .....	94
3.2. EXPERIMENT 1: ENCODING TASK, LONG INTERVAL, 3X .....	95
3.2.1. Methods.....	95
3.2.2. Results .....	99
3.2.3. Discussion .....	103
3.3. EXPERIMENT 2: LURE TYPE, LONG INTERVAL, 3X.....	105
3.3.1. Methods.....	106
3.3.2. Results .....	108
3.3.3. Discussion .....	113
3.4. EXPERIMENT 3: LURE TYPE, SHORT INTERVAL, 3X, ENC. TASK.....	113
3.4.1. Methods.....	114
3.4.2. Results .....	115
3.4.3. Discussion .....	122
3.5. EXPERIMENT 4: LURE TYPE, RETENTION INTERVAL, 6X.....	124
3.5.1. Methods.....	124

3.5.2. Results .....	126
3.5.3. Discussion .....	134
3.6. DISCUSSION OF EXPERIMENTS 1 TO 4.....	138
3.6.1. Empirical summary .....	138
3.6.2. Relation to other length and strength studies .....	140
3.6.3. Implications for memory models.....	142
3.6.4. Limitations .....	145
<b>CHAPTER 4. EXPERIMENTS 5-7 .....</b>	<b>147</b>
4.1. INTRODUCTION .....	147
4.2. EXPERIMENT 5A: RETENTION INTERVAL, WITHOUT NEW, 3X, ONE .....	148
4.2.1. Methods .....	150
4.2.2. Results .....	154
4.2.3. Discussion .....	159
4.3. EXPERIMENT 5B: RETENTION INTERVAL, WITHOUT NEW, 3X, TWO.....	160
4.3.1. Methods.....	161
4.3.2. Results .....	161
4.3.3. Discussion .....	171
4.4. EXPERIMENT 6: RETENTION INTERVAL, WITHOUT NEW, 6X, TWO.....	174
4.4.2. Methods.....	177
4.4.3. Results .....	178
4.4.4. Discussion .....	188
4.5. EXPERIMENT 7: RETENTION INTERVAL, LURE TYPE, WITH NEW, 6X.....	193
4.5.1. Methods .....	195
4.5.2. Results .....	196
4.5.3. Discussion .....	204
4.6. DISCUSSION OF EXPERIMENTS 5 TO 7 .....	207
4.6.1. Empirical summary .....	207
4.6.2. Relation to other experiments .....	210
4.6.3. Implications for memory models.....	214
4.6.4. Limitations .....	234
<b>CHAPTER 5. GENERAL DISCUSSION .....</b>	<b>238</b>
5.1. SUMMARY .....	238

5.1.1. <i>Empirical implications</i> .....	238
5.1.2. <i>Theoretical implications</i> .....	240
5.2. THE ROLE OF RETRIEVAL PRACTICE.....	242
5.3. FURTHER DIRECTIONS .....	244
5.3.1. <i>Cued recall</i> .....	244
5.3.2. <i>Associative recognition</i> .....	249
<b>REFERENCES</b> .....	<b>251</b>
<b>APPENDIX 1</b> .....	<b>269</b>
EXPERIMENT 1 .....	269
EXPERIMENT 2 .....	271
EXPERIMENT 3 .....	272
EXPERIMENT 4 .....	275
EXPERIMENT 5A .....	278
EXPERIMENT 5B .....	281
EXPERIMENT 6 .....	284
EXPERIMENT 7 .....	287
<b>APPENDIX 2</b> .....	<b>290</b>
EXPERIMENT 1 .....	290
EXPERIMENT 2 .....	291
EXPERIMENT 3 .....	291
EXPERIMENT 4 .....	291
EXPERIMENT 5A .....	292
EXPERIMENT 5B .....	293
EXPERIMENT 6 .....	295
EXPERIMENT 7 .....	296
<b>APPENDIX 3</b> .....	<b>298</b>

# List of Figures

FIGURE 1.1. SIGNAL DETECTION INTERPRETATION OF RECOGNITION MEMORY.....	4
FIGURE 1.2. BIND CUE DECIDE MODEL OF EPISODIC MEMORY (BCDMEM). ....	39
FIGURE 1.3. SOUCE OF ACTIVATION CONFUSION (SAC) MODEL. ....	44
FIGURE 1.4. COMPLEMENTARY LEARNING SYSTEMS (CLS) MODEL. ....	50
FIGURE 2.1. UNEQUAL VARIANCE SDT MODEL AND SENSITIVITY MEASURE. ....	67
FIGURE 2.2. EQUAL VARIANCE SDT MODEL AND THE ROC CURVE. ....	69
FIGURE 2.3. EQUAL VARIANCE SDT MODEL WITH FIVE RESPONSE CRITERIA.....	71
FIGURE 2.4. CONSTRUCTION OF ROC CURVE FROM FREQUENCY DATA.....	73
FIGURE 2.5. EFFECT OF CRITERION SHIFTS ON SINGLE-POINT SENSITIVITY. ....	77
FIGURE 2.6. UNEQUAL-VARIANCE SDT MODEL ESTIMATED BY RSCOREPLUS. ....	84
FIGURE 3.1. DESIGN OF EXPERIMENT 1. ....	97
FIGURE 3.2. WORD-FREQUENCY EFFECT ACROSS LIST TYPES (EXP. 1). ....	101
FIGURE 3.3. ROC CURVES FOR EXPERIMENT 1.....	103
FIGURE 3.4. DESIGN OF EXPERIMENT 2. ....	107
FIGURE 3.5. WORD-FREQUENCY EFFECT ACROSS LIST TYPES (EXP. 2). ....	109
FIGURE 3.6. ROC CURVES FOR EXP. 2.....	111
FIGURE 3.7. DESIGN OF EXPERIMENT 3. ....	115
FIGURE 3.8. ROC CURVES FOR EXP. 3.....	119
FIGURE 3.9. SENSITIVITY ACROSS RETENTION INTERVALS (EXPS. 2 / 3). ....	122
FIGURE 3.10. DESIGN OF EXPERIMENT 4. ....	126
FIGURE 3.11. SENSITIVITY ACROSS COMPARISON TYPES (EXP. 4).....	129
FIGURE 3.12. ROC CURVES ACROSS RETENTION INTERVALS (EXP. 4).....	130
FIGURE 3.13. SENSITIVITY ACROSS NUMBER OF REPETITIONS (EXPS. 3 / 4). ....	133
FIGURE 4.1. DESIGN OF EXPERIMENT 5. ....	152
FIGURE 4.2. ROC CURVES FOR A AND B ITEMS (EXP. 5A).....	158
FIGURE 4.3. ROC AND zROC CURVES FOR A AND B ITEMS (EXP. 5B). ....	166
FIGURE 4.4. SENSITIVITY ACROSS NUMBER OF SESSIONS (EXPS. 5A / 5B). ....	169
FIGURE 4.5. DESIGN OF EXPERIMENT 6. ....	177
FIGURE 4.6. ROC CURVES ACROSS RETENTION INTERVALS (EXP. 6).....	184
FIGURE 4.7. SENSITIVITY ACROSS REPETITIONS AND ITEM TYPES (EXPS. 5B / 6)..	187
FIGURE 4.8. DESIGN OF EXPERIMENT 7. ....	195

FIGURE 4.9. SENSITIVITY ACROSS RETENTION INTERVALS (EXP. 7).....	200
FIGURE 4.10. ROC CURVES ACROSS RETENTION INTERVALS (EXP. 7).....	201
FIGURE 4.11. SENSITIVITY ACROSS RETENTION INTERVALS (EXPS. 6-A / 7-SSP).....	204
FIGURE 4.12. SENSITIVITY DATA FROM FIRST STUDY-TEST BLOCK (EXP. 7). .....	236



## List of Tables

TABLE 3.1. HITS AND FALSE ALARMS (EXP. 1). .....	100
TABLE 3.2. SENSITIVITY ( $A_z$ ) AND BIAS ( $C_A$ ) (EXP. 1). .....	102
TABLE 3.3. HITS AND FALSE ALARMS (EXP. 2). .....	108
TABLE 3.4. SENSITIVITY ( $A_z$ ) ACROSS DISCRIMINATION TYPES (EXP. 2). .....	110
TABLE 3.5. BIAS ( $C_A$ ) ACROSS DISCRIMINATION TYPES (EXP. 2). .....	112
TABLE 3.6. HITS AND FALSE ALARMS ACROSS ENCODING TASKS (EXP. 3). .....	116
TABLE 3.7. SENSITIVITY ( $A_z$ ) ACROSS DISCRIMINATION TYPES (EXP. 3). .....	118
TABLE 3.8. BIAS ( $C_A$ ) ACROSS DISCRIMINATION TYPES (EXP. 3). .....	121
TABLE 3.9. HITS AND FALSE ALARMS (EXP. 4). .....	127
TABLE 3.10. SENSITIVITY ( $A_z$ ) ACROSS DISCRIMINATION TYPES (EXP. 4). .....	128
TABLE 3.11. BIAS ( $C_A$ ) ACROSS DISCRIMINATION TYPES (EXP. 4). .....	132
TABLE 4.1. HITS AND FALSE ALARMS ACROSS ITEM TYPES (EXP. 5A). .....	155
TABLE 4.2. SENSITIVITY ( $A_z$ ; SSP COMPARISON) ACROSS ITEM TYPES (EXP. 5A). .....	157
TABLE 4.3. BIAS ACROSS ITEM TYPES ( $C_A$ ) (EXP. 5A). .....	159
TABLE 4.4. HITS AND FALSE ALARMS ACROSS ITEM TYPES (EXP. 5B). .....	162
TABLE 4.5. SENSITIVITY ( $A_z$ ; SSP COMPARISON) ACROSS ITEM TYPES (EXP. 5B). .....	165
TABLE 4.6. BIAS ACROSS ITEM TYPES ( $C_A$ ) (EXP. 5B). .....	167
TABLE 4.7. HITS AND FALSE ALARMS ACROSS ITEM TYPES (EXP. 6). .....	181
TABLE 4.8. SENSITIVITY ( $A_z$ ; SSP COMPARISON) ACROSS ITEM TYPES (EXP. 6). ..	183
TABLE 4.9. BIAS ACROSS ITEM TYPES ( $C_A$ ) (EXP. 6). .....	185
TABLE 4.10. SENSITIVITY ( $D'$ ) IN CARY AND REDER (2003). .....	189
TABLE 4.11. HITS AND FALSE ALARMS IN NORMAN (1999). .....	190
TABLE 4.12. HITS AND FALSE ALARMS (EXP. 7). .....	197
TABLE 4.13. SENSITIVITY ( $A_z$ ) ACROSS RETENTION INTERVALS (EXP. 7). .....	199
TABLE 4.14. BIAS ( $C_A$ ) ACROSS DISCRIMINATION TYPES (EXP. 7). .....	203
TABLE 4.15. EFFECT SIZES OF LLEs AND LSEs (EXP. 7). .....	205
TABLE 4.16. WITHIN-LIST, STRENGTH-BASED MIRROR EFFECT (EXPS. 5 / 6). ....	229

# List of Abbreviations

<b>LLE</b>	List-Length Effect
<b>LSE</b>	List-Strength Effect
<b>Short</b>	Short list (all weak items; presented once)
<b>Long</b>	Long list (all weak items)
<b>Strong</b>	Strong list (half weak items; half strong items)
<b><i>H</i></b>	Hit rate (proportion of correct “old” responses)
<b><i>F</i></b>	False-alarm rate (proportion of incorrect “old” responses)
<b><math>\mu</math></b>	Mean familiarity for target ( $\mu_T$ ) and distractor ( $\mu_D$ ) distributions
<b><math>\sigma</math></b>	Standard deviation for target ( $\sigma_T$ ) and distractor ( $\sigma_D$ ) distributions
<b><i>M</i></b>	Sample mean
<b><i>SD</i></b>	Sample standard deviation
<b><i>N</i></b>	Number of participants
<b><i>SEM</i></b>	Standard error of the mean ( $= SD / \sqrt{N}$ )
<b><i>d'</i></b>	Discriminability (when distributions have equal variance)
<b><i>d<sub>a</sub></i></b>	Discriminability (when distributions have unequal variance)
<b><i>X<sub>i</sub></i></b>	Decision criterion ( $i = 1$ to 6)
<b><i>c</i></b>	Decision criterion (equal variance)
<b><i>c<sub>a</sub></i></b>	Decision criterion (unequal variance)
<b><math>\Phi(X)</math></b>	Cumulative normal distribution ( $X$ = value of random variable)
<b><math>z(P)</math></b>	Inverse cumulative normal distribution ( $P$ = proportion)
<b>ROC</b>	Receiver Operating Characteristic curve ( $y$ = hits, $x$ = false alarms)
<b><i>z</i>ROC</b>	Normalised ROC [ $y = z(H)$ , $x = z(F)$ ]
<b><i>A<sub>z</sub></i></b>	Area under ROC, normal distribution assumed
<b>BCDMEM</b>	Bind Cue Decide Model of Episodic Memory
<b>CLS</b>	Complementary Learning Systems model
<b>SAC</b>	Source of Activation Confusion model
<b>SAM</b>	Search of Associative Memory model
<b>REM</b>	Retrieving Effectively from Memory model
<b>TODAM</b>	Theory of Distributed Associative Memory

## Declaration

This thesis is the work of the author and has not been submitted for a degree at another university. Portions of Chapter 3 (Experiment 3 in section 3.4) have been published in Buratto, L. & Lamberts, K. (2008). List-strength effect without list-length effect in recognition memory. *Quarterly Journal of Experimental Psychology*, 61, 218-226.

# Acknowledgments

There are many people who have helped me immensely at different stages of this project. I would like to thank Nick Chater for bringing me to Warwick and Koen Lamberts for “adopting” me after Nick left for University College London. Koen has been a constant source of support and wise advice throughout this period and I consider myself lucky to have inadvertently ended up under his supervision.

I would like to thank Menelaos Apostolou, Elizabeth Blagrove, Duncan Guest, William Jimenez-Leal, Chris Kent, William Matthews, Kerry McColgan, Joanne Myers, Erika Nurmsoo, Maria Sapouna, Christoph Ungemach, and Theodora Zarkadi for making my stay at Warwick a much more enjoyable experience.

I also would like to thank the British government and Warwick University’s Department of Psychology for providing me with funding (Overseas Research Students Award Scheme and Warwick Postgraduate Research Fellowship, respectively), which allowed me to work during these three-and-a-half years without financial worries.

Finally, I am greatly indebted to my parents, Paulo and Julieta, who always encouraged me to study (in good faith, I would say, as they have only a vague idea of what I am up to!), my brother, Gabriel, for his friendship and for providing me with constant motivation for spelling out clearly any ideas I might have, and my wife, Noeli, who has been by my side through it all, for her patience, beauty and love. To her I dedicate this thesis.

# Abstract

The study of interference effects is important to constrain models of memory. List-length manipulations test how adding new information to memory affects memory for the other stored information (list-length effect; LLE). List-strength manipulations test how strengthening some information in memory affects memory for the other non-strengthened information (list-strength effect; LSE). Whereas LLE and LSE are generally found in recall tasks, their empirical status in recognition tasks is less well established. In this thesis, we investigated some boundary conditions for both list-length and list-strength effects. The results provided evidence for the following claims: *i*) LLE and LSE are real effects in recognition (the effects were obtained after controlling for several confounds); *ii*) LLE and LSE are modulated by the relative contribution of recall-like processes operating at test (more recollection at test yielded larger effects); *iii*) LLE and LSE can be modulated by the number of study-test blocks in an experimental session (fewer study-test blocks resulted in larger effects); *iv*) LLE and LSE can be modulated by the time interval between study and test (shorter intervals produced larger effects) and *iv*) LLE and LSE may not be strongly modulated by the magnitude of length and strength manipulations (stronger manipulations did not result in larger effects). Taken together, the results support memory models that attribute forgetting in recognition to competition between memory traces during either encoding or retrieval. The results provide little support for models that attribute forgetting solely to interference between the contexts in which a memory was originally stored.

# Chapter 1. Review and objectives

How is education supposed to make me feel smarter? Every time I learn something new, it pushes some old stuff out of my brain. Remember when I took that home winemaking course, and I forgot how to drive? **Homer Simpson**

## 1.1. Introduction

The study of interference is an important element of memory research. Is it the case that the more one learns, the more one forgets? Is it the case that learning something well comes at the cost of forgetting something else? In this thesis, we focus on how memories interact when people undergo recognition tests, where the task is to distinguish whether or not a given piece of information has been previously seen. In particular, we are interested in what happens to the memory of a given piece of information when many new pieces of information are learned once or when few new pieces of information are learned over and over again.

In this type of research, “piece of information” is usually represented by a “word” and the “contents of memory” are represented by a “list of words”. Two manipulations have been commonly used to assess interference in memory: list-length and list-strength manipulations. List-length manipulations test how adding items to a list of words affects memory for the other words on the list. A list-length effect (LLE) involves better performance on short lists than on long lists. List-strength manipulations, on the other hand, test how strengthening items (e.g., by repetition or study time) affects memory for the other, non-strengthened items on the list. A list-strength effect (LSE) occurs when performance on non-strengthened items is better in *pure weak* lists (where all items have the same strength) than in *mixed* lists (where some items have been strengthened).

Length and strength manipulations are *empirically* interesting because they allow us to evaluate how stored items affect each other during memory tasks. Those manipulations are also *theoretically* important because they test core assumptions of several computational memory models. Indeed, a whole class of models – the global matching models – was developed to explain, among other things, how

items stored in memory interfere with each other during storage or retrieval. This type of interference could potentially explain why forgetting occurs.

Although global matching models predict LLE and LSE, and although the existence of such effects could be almost taken for granted (i.e., it is intuitive to think that adding items to a list should impair the memory of any one item), empirical results in recognition studies have been mixed. Both positive and null LLEs and positive and null LSEs have been reported. Those findings represent a challenge to established memory models. Indeed, new memory models have been developed in recent years trying to explain the mixed pattern of results.

Given the uncertain status of length and strength effects in recognition and given their theoretical importance, it is crucial to carefully investigate the boundary conditions underlying those effects. The main aim of this thesis is to present evidence bearing on such boundary conditions. In this chapter, we describe why length and strength manipulations are theoretically important. Next, we introduce the list-length and list-strength paradigms (and their variations) and review the main empirical findings, discussing some limitations of previous research.

## **1.2. Early global matching memory models**

In the 1980s, several process models were developed that were able to account for findings in a wide range of experimental paradigms, including categorisation, recall and recognition (hence the name *global models*). We will focus on the three most investigated models from that generation: Search of Associative Memory (SAM; Gillund & Shiffrin, 1984), MINERVA2 (Hintzman, 1988) and TODAM (Theory of Distributed Associative Memory; Murdock, 1982).

Although differing in many respects, those models share two common assumptions when applied to recognition memory: they assume that all items stored in memory contribute information to the recognition decision and that the information contributed by each item in memory is evaluated in parallel. In other words, when an item is presented at test, the information used to assess whether or not that item has been previously seen contains simultaneous contributions

from all items stored in memory. This information signal can be interpreted as an index of the *match* between the test items and the contents of memory (hence the name *matching models*), or alternatively, as an index of the *familiarity* of the test item. Thus the more familiar a test item, the higher the familiarity signal.

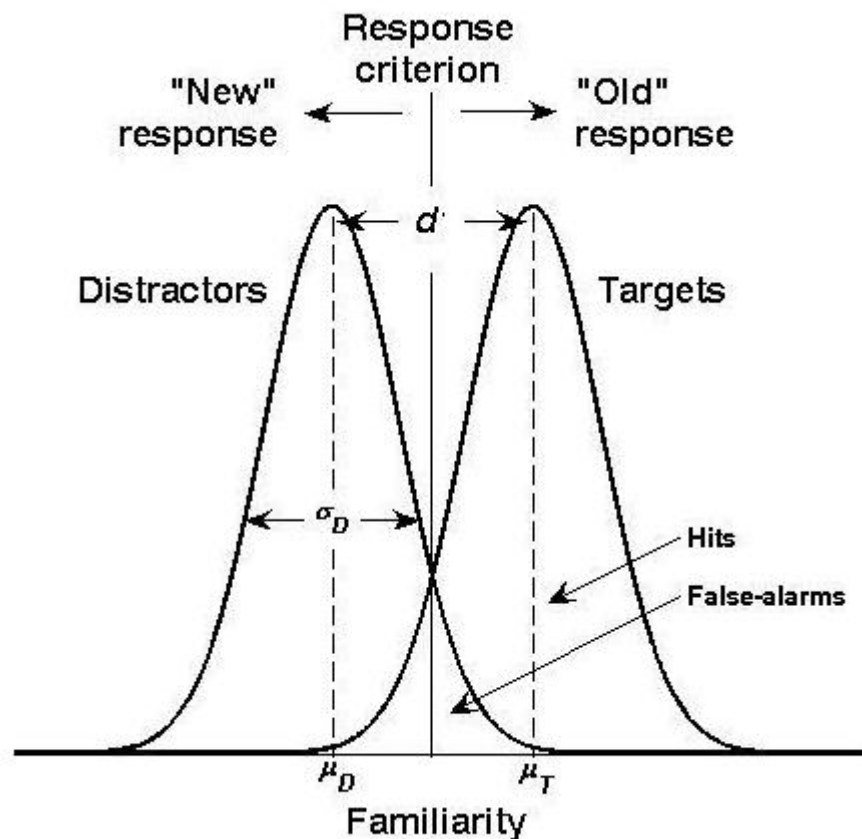
The matching assumption allows SAM, MINERVA2 and TODAM to explain phenomena that were difficult for previous models to account for, namely, the speed of recognition decisions and similarity effects. High confidence memory decisions are made fast (Glucksberg & McCloskey, 1981); matching models can account for that through the parallelism of the matching process. Matching models can also account for similarity effects – the finding that false recognitions are high when the items stored in memory are similar to each other and to the test item (e.g., Posner & Keele, 1970) – because they take into account all stored items during the recognition process.

The familiarity signal produced by the matching process can then be analysed within the framework of Signal Detection Theory (SDT; Macmillan & Creelman, 2005). The familiarity of a given class of items (e.g., high-frequency words, concrete words) is assumed to follow a normal distribution with mean  $\mu$  and standard deviation  $\sigma$  (a more detailed description of SDT applied to recognition memory is given in Chapter 2). The familiarity distribution of studied items (also called *target* or *old* items) is assumed to have higher mean and standard deviation than the distribution of unstudied items (also called *distractor*, *new*, *foil* or *lure* items). Performance at this level of analysis is thus a function of the means and standard deviations of *target* ( $\mu_T, \sigma_T$ ) and *distractor* ( $\mu_D, \sigma_D$ ) distributions. Global matching models are able to produce estimates of means and standard deviations, and those estimates are used to predict recognition performance. One commonly used measure of discriminability is given by:

$$d' = \frac{\mu_T - \mu_D}{\sigma_T} \quad (1.1)$$

In other words, the ability to discriminate studied from unstudied items is proportional to the difference between the means of *targets* and *lures* and inversely proportional to the standard deviation of the *target* distribution. Figure 1.1 illustrates the recognition decision according to the SDT framework.





**Figure 1.1. Signal detection interpretation of recognition memory.**

The decision axis represents the familiarity scale, ranging from low to high. The two Gaussian distributions represent familiarities associated with targets (mean  $\mu_T$ ) and distractors (mean  $\mu_D$ ; standard deviation  $\sigma_D$ ). Standard deviations are assumed to be equal. Discriminability  $d'$  is the standardised difference between the means of targets and distractors. The vertical bar (*criterion*) separates the decision space between “old” responses (i.e., “I have seen the item”) and “new” responses (i.e., “I have not seen the item”). Hits represent the proportion of “old” responses given to targets; false alarms represent the proportion of “old” responses given to distractors.

Because global matching models take into account information about other items in memory during the matching process, it is possible that the memory of a given target item is impaired by manipulations affecting the remaining items in memory (e.g., adding items to memory or strengthening some items in memory). SAM, MINERVA2 and TODAM all predict a decrease in discriminability ( $d'$ ) as a result of list-length and list-strength manipulations. In the following, we briefly describe those models and explain why they predict LLE and LSE (for a review of global matching models, see Clark & Gronlund, 1996).

### 1.2.1. SAM (Gillund & Shiffrin, 1984)

In this model, memory is represented as a matrix of association strengths between memory traces and the cues used to activate those traces. Associations are generated during learning through a rehearsal process modulated by four parameters:  $a$  modulates the association strength between study context and study item (contextual association);  $b$  modulates the association between items that were studied together (inter-item association);  $c$  modulates the association between a studied item and its own trace; and  $d$  modulates the association between a studied and a non-studied item (i.e., pre-experimental association).

Because traces in SAM are stored separately from each other, interference does not occur at storage. Instead, forgetting is due to events unfolding at the time of retrieval. During the retrieval process, a familiarity signal is produced through the activation of all stored memories by the cues available at test (i.e., the test context and the test item). If  $C$  represents the test context cue and  $I$  represents the item cue, then the familiarity signal elicited by a test item  $I$  is given by

$$F(C, I) = \sum_{j=1}^N S(C, I_j)^{W_C} \times S(I, I_j)^{W_I} \quad (1.2)$$

where  $N$  is the number of traces in memory,  $W_C$  and  $W_I$  are attention weights (e.g.,  $W_C = W_I = 0.5$ ) and  $S(X, Y)$  are the association strengths between cue  $X$  and memory trace  $Y$ . To illustrate, if the cue  $C$  is the test context, then  $S(C, I_j) = a$ , and if cue  $I$  represents trace  $I_j$  stored in memory, then  $S(I, I_j) = c$ . Equation 1.2 implements the matching assumption, since familiarity is obtained by summing over all traces stored in memory. To produce a response, the model compares the familiarity signal to a decision criterion: if the signal is higher than the criterion, then an “old” response is produced. Otherwise, a “new” response is produced.

In order to avoid perfect performance, variability is added to the association strengths  $a$ ,  $b$ ,  $c$  and  $d$ . The assumption adopted by Gillund and Shiffrin (1984) involved replacing each strength  $X$  in the memory matrix with a value taken from a 3-point uniform distribution given by  $0.5X$  or  $X$  or  $1.5X$ . As a result, the means and variances of associative strength distributions are tied in the SAM model: the higher the mean strengths, the higher the variance of their distributions.

List-length and list-strength effects follow from SAM's variability assumptions. As shown in Equation 1.1, discriminability is a function of two factors: the difference in the mean familiarity values of targets and distractors (numerator); the size of the distractor variance (denominator). The numerator is not affected by length or strength: adding items or strengthening some items increments the mean familiarity of both targets and distractors by the same amount (see Clark & Gronlund, 1996, p. 41, for an example). The denominator, however, is predicted to increase. The terms in the summation in Equation 1.2 are independent (by assumption) and it is known from probability theory that the variance of the sum of independent variables is the sum of their variances. Thus adding more items to a list increases both target and distractor variances (more  $S(C, I_j) \times S(I, I_j) = ad$  terms are added to the familiarity of both distributions). Similarly, strengthening some items increases the values of parameter  $a$  for those items (the association between strong items and context is higher than the association between weak items and context), and higher  $a$  implies higher variance. The increase in variability (denominator in Equation 1.1) causes performance ( $d'$ ) to decrease. In sum, the SAM model predicts both list-length and list-strength effects in recognition, and these predictions follow from core assumptions of the model.

### 1.2.2. MINERVA2 (Hintzman, 1988)

Memory traces are represented as separate vectors containing  $M$  features. Each feature can assume values +1, 0 or -1. Features of study items are correctly stored with probability  $L$  (0 is stored with probability  $1 - L$ ). Variability in the matching process comes from the probabilistic nature of feature encoding. The familiarity signal produced by a test item is obtained by taking the dot product of the test item with each vector stored in memory. If  $T$  represents a stored vector and  $P$  represents the test vector, then the familiarity elicited by  $P$  at test is given by

$$F(P) = \sum_{i=1}^N \left( \sum_{j=1}^M \frac{T_{ij} P_j}{N_{\pm, i}} \right)^3 \quad (1.3)$$

where  $N$  is the number of traces in memory,  $T_{ij}$  is the value of feature  $j$  in trace  $i$ ,  $P_j$  is the value of feature  $j$  in the test item and  $N_{\pm, i}$  is the number of non-zero features in both vectors. The cubic exponent in each term of Equation 1.3 allows

the model to boost the signal produced by test cues that are similar to stored traces and shrink the signal from cues that are dissimilar to stored traces. The familiarity signal is then compared to a criterion in order to output a response.

Like SAM, MINERVA2 predicts the occurrence of list-length and list-strength effects at retrieval due to an increase in the variance of familiarity distributions. The mean difference between targets and distractors does not change because, for each non-target trace activated by an old test item, there is an equivalent activation of that trace by a new test item. By contrast, the variance increases with longer and stronger lists because each stored trace is an independent and additive source of variance to the familiarity signal (the encoding probability  $L$  is applied independently to each feature of each trace). Thus MINERVA2 also predicts both list-length and list-strength effects in recognition.

### 1.2.3. TODAM (Murdock, 1982)

In TODAM, each item is represented by a vector of  $N$  features. Features are either encoded (probability  $p$ ) or not encoded (set to 0 with probability  $1 - p$ ). Each feature value is chosen from a distribution with mean 0 and variance  $1/N$ . All items are stored in a common vector. Thus, unlike SAM and MINERVA2, where traces are stored separately, representation in TODAM is composite. Consequently, forgetting is assumed to occur at storage.

If  $M_{i-1}$  is the composite vector containing  $i - 1$  traces and  $f_i$  is a new trace, then the updated version of  $M_{i-1}$  is given by  $M_i = \alpha M_{i-1} + p f_i$ , where  $\alpha$  ( $< 1$ ) is a forgetting parameter (i.e., before encoding  $f_i$ ,  $M$  is decremented by  $\alpha$ ). The familiarity signal associated with test item  $g$  is produced in TODAM by taking the dot product between test item  $g$  and the memory vector  $M$ , such that:

$$F(g) = g \cdot M = \sum_{i=1}^N g_i M_i \quad (1.4)$$

If  $g$  matches a trace in memory (i.e., a vector that was added to the composite vector  $M$ ), then the mean familiarity of that match is  $p\alpha^{L-i}$ , where  $L$  is the list length and  $i$  is the serial position where item  $g$  was studied. If the test item does not match any trace encoded in  $M$ , then the mean familiarity is 0. The variance of a match is given by  $2p/N$  and the variance of a mismatch is given by  $p/N$  (see

Weber, 1988, Table 1). Thus higher encoding probability  $p$  (obtained by item repetition, for example) entails higher variance. Like SAM and MINERVA2, the outcome of the matching process is compared to a decision criterion in order to emit an old-new response.

TODAM predicts LLE because the familiarity of each item is decreased for every new item added to the list (i.e.,  $\text{match} = p\alpha^{L-i}$ ). Moreover, every extra item adds another source of variance. Thus the numerator of Equation 1.1 decreases and the denominator increases, resulting in a decrease in  $d'$ . TODAM predicts LSE because stronger lists have higher mean variances than weaker lists.

### 1.3. Empirical evidence: early studies

The global matching models discussed in the previous section predict list-length and list-strength effects in recognition. But what is the evidence supporting those predictions? Here we review some early studies suggesting that LLE is a real phenomenon in recognition whereas LSE is not.

#### 1.3.1. Evidence for list-length effect

List-length effect has long been observed in recall (e.g., Ebbinghaus, 1885/1964; Murdock, 1962) and recognition (e.g., Strong, 1912) and has thus been treated as a standard phenomenon to be explained by any memory model.

In his seminal study, Strong (1912) asked participants to study sequences of full-page advertisements. The sequences contained 5 to 150 advertisements. At test, old and new items were presented and participants had to sort each advertisement into piles according to confidence. The results showed that *hit rates* (proportion of correct “old” responses) decreased with list length and *false alarm rates* (proportion of incorrect “old” responses) increased with list length.

Strong’s (1912) study, however, confounded list length with study-test lag.<sup>1</sup> The test was carried out immediately after study in both the short and long conditions.

---

<sup>1</sup> *Study-test lag* is the time interval between the study of an item and its subsequent test. This is not to be confused with *retention interval*, which refers to the time interval between the end of the study list and the beginning of the test list. We follow this convention throughout this thesis.

As a consequence, the average study-test lag per item was shorter in the short list than in the long list. Because recognition drops with longer lags (e.g., Shepard, 1967; Strong, 1913), either list length or study-test lag could have accounted for the results. Another confound refers to the size of the test list. Longer study lists were followed by longer test lists. But it is known that discrimination on the latter part of a long list is impaired relative to discrimination at the beginning of the list (Schulman, 1974). Thus average performance in a long test list is worse than in a short test list, regardless of the length of the study list. Finally, Strong's (1912) study did not control for serial position effects. Items at the beginning and at the end of a list are better recognised than items in the middle (Neath, 1993; Schulman, 1974). This primacy and recency advantage would benefit short lists more than long lists because a higher proportion of items in the short list would partake of the gain.

Several subsequent studies continued to confound list length with other variables (e.g., length of test list in Murnane & Shiffrin, 1991a; study-test lag in Yonelinas, 1994). But the effect was still observed even when most of those confounds were eliminated. Gronlund and Elam (1994), for example, found a strong and reliable LLE when using a *retroactive design*. This experimental design addressed early criticisms because it equated study-test lag between short and long conditions (the period after the presentation of study items was filled with a distractor task) and because it controlled for serial position effects (only items studied at the beginning of both lists were compared).<sup>2</sup>

Despite the apparent reality of LLE in recognition, Murdock and Kahana (1993a, 1993b) argued that the effect should disappear when the number of items intervening between study of an item and its test is controlled. This prediction was derived from a modified version of TODAM proposed to account for the absence of LSE in recognition (see 1.3.2 and 1.4.2). In most studies, targets on the study list and targets and distractors on the test list are randomly mixed. Thus it is usually not possible to assess directly the effect of the number of intervening

---

<sup>2</sup> The fact that only early items in both study lists enter the analysis suggests that any ensuing memory impairment is likely to have been caused by the later study items. In this sense, the retroactive design used in recognition is similar to the retroactive interference paradigm used in cued recall (study AB and AC pairs; test AB pairs only: Barnes & Underwood, 1959).

items on memory. To address this issue, Ohrt and Gronlund (1999, Exp. 1) conducted an experiment using a *proactive design*. In this design, study-test lag is equated (the period after the end of both lists is filled with a distractor task), serial position effects are controlled (only items studied at the end of both lists are compared) and the number of intervening items is taken into account (items studied early are tested early; items studied late are tested late).<sup>3</sup> Unlike the retroactive design, where the distractor task is longer in the short list than in the long list, the proactive design requires equal distractor times for both list lengths. Ohrt and Gronlund (1999) found a significant LLE. The effect was replicated in a second experiment where category length was manipulated and study-test positions were controlled. These results, together with the results of similar studies (e.g., Ratcliff, McKoon, & Tindall, 1994, Exp. 3), suggest that the LLE is not an experimental artifact; apparently it represents a real phenomenon.

### 1.3.2. Evidence against list-strength effect

Unlike list-length manipulations, which have been studied since Ebbinghaus in the late 1800s, list-strength manipulations have attracted attention only during the last 30 years. Tulving and Hastie (1972), the first to investigate the issue, found an LSE in free recall (see also Malmberg & Shiffrin, 2005; Ratcliff, Clark, & Shiffrin, 1990, Exp. 6; Rose & Sutton, 1996; Wixted, Ghadisha, & Vera, 1997). Most studies up until the end of the 1990s, however, have found only a small LSE in cued recall (Ratcliff, Clark, & Shiffrin, 1990, Exps. 3 and 6) and no effect in recognition at all (Hirshman, 1995; Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990; Ratcliff, Gronlund, & Sheu, 1992; Ratcliff et al., 1994; Yonelinas, Murdock, & Hockley, 1992).

Strength manipulations have usually been implemented using the *mixed-pure paradigm* (Ratcliff et al., 1990). Participants are presented with *mixed lists* containing weak items (e.g., presented once or for a short time) and strong items (e.g., presented more than once or for a longer time). As the goal is to assess the impact strengthening some items has on non-strengthened items, it is necessary

---

<sup>3</sup> As only late items in both study lists enter the analysis, any forgetting is likely to have been caused by earlier study items. In this sense, the proactive design is similar to the proactive interference paradigm in cued recall (study AB and AC pairs; test AC only: Underwood, 1957).

to create appropriate controls against which to compare the performance of weak and strong items in the mixed lists. There are two possible controls: *pure weak lists* are lists containing only weak items and *pure strong lists* contain only strong items. The number of unique items in each list is held constant to control for list-length effects. A list-strength effect occurs if: *i*) weak item discrimination is better in *pure weak* lists than in *mixed* lists; *ii*) strong item discrimination is better in *mixed* lists than in *pure strong* lists.

Ratcliff et al. (1990) conducted 7 experiments, none of which showed the predicted differences between pure and mixed lists. Moreover, in their Experiment 6, list length was manipulated along with list strength and the results yielded a dissociation whereby LLE was observed and LSE was not. Null results are difficult to interpret in general (Frick, 1996) and because this particular result (null LSE) has serious implications for global matching models, it is important to make every possible effort to rule out confounds in the experimental design.

One important confound in designs using mixed lists is *rehearsal redistribution*, which occurs when participants take effort or rehearsal time away from strong items in a mixed list and redistribute that effort or time to weak items in the same list. This may occur because after a few presentations of the same word, participants may feel they already know the item well enough and that they should spend some time practicing other less-well-learned items. To the extent that this rehearsal redistribution occurs, it works against the possibility of finding an LSE, as weak items would receive more rehearsal time (and become stronger) and strong items would receive less rehearsal time (and become weaker). Therefore, any differences between weak and strong items across pure and mixed lists would be reduced and could potentially disappear. Although rehearsal redistribution can occur in pure lists (either weak or strong), it does not work against finding an LSE in those lists because any rehearsal taken away from an item is reallocated to another item of the same class.

Several studies converged on the conclusion that rehearsal redistribution was not the reason behind the null LSE. For example, Ratcliff et al. (1990) blocked weak and strong items (such that if redistribution did occur, it would likely have



occurred at the block boundaries); Murnane and Shiffrin (1991b) analysed their data as a function of the strength of a given item's neighbours (assuming that redistribution occurs mainly between adjacent items); Yonelinas et al. (1992) presented words for very short periods of time (50 ms for weak items; 200 ms for strong items), under the assumption that rehearsal strategies take time and that during such short presentation times participants would not have enough time to read the word and simultaneously adopt a redistribution strategy. In all these studies, where rehearsal redistribution was unlikely to operate, no LSE emerged.<sup>4</sup>

The experimental confounds present in early list-length experiments (see 1.3.1) were also present in early list-strength experiments (e.g., study-test lag in Yonelinas et al., 1992). Stronger lists are also longer in total presentation time. Therefore, a study-test lag confound should work towards finding an LSE because weak items in weak lists would have a shorter study-test lag, on average, than weak items in mixed lists and strong items in pure lists would have a longer study-test lag than strong items in mixed lists. The fact that an LSE was not found even in experiments containing study-test lag confounds can be interpreted as support for the null finding. In any case, when study-test lag was controlled for, the results unsurprisingly showed no sign of an LSE (e.g., Murnane & Shiffrin, 1991a, Exps. 3 and 6).

The only visible effect of the list-strength manipulation on mixed-list items was observed in the setting of the *decision criterion*. The criterion corresponds to the cut-off value in familiarity space that separates “old” from “new” responses. It represents the minimum degree of familiarity elicited by a test item that a participant deems sufficient to emit an “old” response. Hirshman (1995) found that criterion placement varied systematically with list strength: it increased from *pure weak* to *mixed* to *pure strong* lists, indicating that participants became more and more conservative in their output of “old” responses as average item strength

---

<sup>4</sup> Murnane and Shiffrin (1991b) found an LSE only when items were likely to be stored separately in memory by using sentences as study items and by repeating the words in the sentences in the context of different sentences (as if repeated items were new items in a list-length procedure). When items were unlikely to be stored separately, by repeating the same sentences, the LSE disappeared. Yonelinas et al. (1992) also found an LSE but the effect was later attributed to reverse rehearsal redistribution: participants were redistributing effort from weak to strong items, presumably because the weak items were presented too quickly (50 ms) to be worth attending to (i.e., focusing on the slower strong items would yield better performance).

increased. Hirshman (1995) found this pattern not only in his experiments but also in most previously published list-strength studies (the pattern held in 75 out of 92 comparisons between conditions).

Criterion can be estimated by observing the behaviour of hits and false alarms across conditions. If hits and false alarms move in the same direction (i.e., they increase or decrease in tandem), then one may claim a criterion shift occurred.<sup>5</sup> Criterion setting is assumed to be under the participants' strategic control. But because no instructions expected to affect criterion setting are usually given to participants in list-strength experiments, criterion has been treated by global matching models simply as a parameter to be estimated, without any real mechanism accounting for its behaviour. For our present purposes, it suffices to say that, unlike list-length manipulations, where criterion setting rarely changes between conditions (hits and false alarms move in opposite directions), list-strength manipulations tend to cause criterion shifts across pure and mixed list conditions (hits and false alarms decrease in tandem).

Because list-strength manipulations affect criterion setting across conditions, and because criterion setting may have an effect in discriminability measures such as  $d'$  (see Van Zandt, 2000, for recent evidence), it is important to control criterion placement in any demonstration of a null LSE. To address this issue, Shiffrin, Huber and Marinelli (1995) used long lists of categorised items. Participants are reluctant to change the decision criterion within the same list, especially when no feedback is provided after each trial (Stretch & Wixted, 1998b; Verde & Rotello, 2007). Length was manipulated by increasing the number of items in a category and strength was manipulated by increasing the number of presentations of an item in a category. As expected, criterion did not change across conditions (false alarms for weak and strong items were the same for pure and mixed categories). More importantly, the results revealed a positive LLE but no LSE.

In short, the results of several studies carried out up until the end of the 1990s were consistent with the idea that there is no list-strength effect in recognition.

---

<sup>5</sup> Unidirectional changes in hits and false alarms can also be caused by a shift in the underlying familiarity distributions (e.g., old and new distributions move up) without any change in criterion.

Those results represented a challenge to global matching models, since they all predicted the existence of an LSE directly from their basic assumptions.

### 1.3.3. Other challenges to global matching models

The null LSE was not the only challenge faced by global matching models at that time. Other results also started to cast doubt on some of the models assumptions. Here we discuss a series of results that went counter to the predictions made by the global models SAM, MINERVA2 and TODAM.

As discussed in 1.2, global matching models predict that the variance of old and new items should increase with increases in list length and list strength. In particular, SAM and MINERVA2 predicted that item strength should increase the variance of the old-item distribution more than the variance of the new-item distribution. That should occur because the match value of an old test item depends on how strongly that item was encoded in memory: the higher the strength in memory, the higher the match value.

An old test item matches one trace in memory and mismatches all the other traces, whereas a new test item mismatches all traces in memory. When the match value of an old item increases, the contribution to the overall old-item variance starts to be dominated by the value of the strong match over the values of all the mismatches. The variance of new-item distribution, on the other hand, contains only the contribution of mismatches. Therefore, the variance of old items should increase relative to the variance of new items when strength is manipulated. In TODAM, by contrast, the variances of old and new distributions should not appreciably change, as the variance of a match is about twice the variance of a mismatch, regardless of strength. Thus the new-to-old variance ratio should either decrease with list strength (ratio less than 1, according to SAM and MINERVA2) or remain constant (ratio equal to 1, according to TODAM). Ratcliff, Gronlund and Sheu (1992) found a constant variance ratio, disconfirming the predictions of the three models.<sup>6</sup>

---

<sup>6</sup> Ratcliff et al.'s (1992) estimate of the new-to-old standard deviation ratio was 0.8. In Chapter 2, we describe how standard deviation ratios ( $s = \sigma_D/\sigma_T$ ) and, consequently, variance ratios can be estimated from participants' hits and false alarms.

Subsequent studies confirmed the finding and extended it to list-length manipulations. When list length increases, the number of traces encoded in memory also increases. This leads to an increase in the variance of old and new item distributions as there are more independent terms added to the familiarity sum. As the list length increases, the contribution of mismatches to the overall variance of the old distribution starts to dominate the contribution of matches. Similarly, the new-item variance is made up exclusively of mismatches. Thus, as list length increases, the new-to-old variance ratio should approach 1. This prediction is shared by SAM, MINERVA2 and TODAM. Contrary to the prediction, Gronlund and Elam (1994, Exp. 1) and Ratcliff et al. (1994, Exp. 3) found that the variance ratio was constant across list lengths and set around 1.

The fact that the estimated new-to-old variance ratios were constant across list-length and list-strength manipulations contradicted core assumptions of the global matching models. Because those assumptions underlay most of the published predictions for those models, major revisions became necessary.

## **1.4. Recent global matching memory models**

The challenges presented to global matching models following their failure to predict several patterns of results led to changes in some of their basic assumptions and led to the development of a new generation of matching models, including REM (Shiffrin & Steyvers, 1997) and SLiM (McClelland & Chappell, 1998). In the following, we describe the modifications implemented in the early matching models to account for the null LSE, how those modifications were incorporated into REM and how they affected the predictions for LLE and LSE.

### **1.4.1. SAM and the differentiation assumption**

In SAM, the strength parameter  $d$  representing the association between a lure test item and a trace in memory (i.e., a form of pre-experimental associative strength) is assumed to be constant, regardless of trace strength. The parameter  $a$  (contextual strength), on the other hand, increases with trace strength. Thus larger  $ad$  terms are added to the familiarity signal of new test items as old-item

strength is increased. If instead  $d$  is assumed to decrease when the strength of the memory traces increases, then a form of *differentiation* has been implemented: as a trace becomes stronger, it becomes more connected to the study context (increase in  $a$ ) and less connected to unstudied items (decrease in  $d$ ). In other words, strong items become increasingly distinct from weak items. And because the new-item variance remains largely unchanged, an LSE is not predicted (see Shiffrin, Ratcliff, & Clark, 1990, for a detailed discussion). Conversely, an LLE is still predicted because the differentiation assumption does not apply for list-length manipulations, as all added items are equally strong.

Although the differentiation assumption allows SAM correctly to predict a null LSE, it does so through a careful balance between context ( $a$ ) and residual ( $d$ ) strengths. Positive, null or negative LSE can be predicted depending on the relationship between trace strength and  $d$  and on the relative weights assigned to context ( $W_C$ ) and item ( $W_I$ ) cues (see Shiffrin et al., 1990, Fig. 1). Moreover, although the differentiation assumption correctly predicts an LLE, it still incorrectly predicts that the new-to-old variance ratio should approach 1 with increasing length (Gronlund & Elam, 1994). Thus, despite the differentiation assumption's ability to fix some of the problems faced by global matching models, it is not able alone to account for the whole pattern of empirical data.

#### 1.4.2. TODAM and the continuous memory assumption

In TODAM, the composite vector  $M$  containing the traces of all studied items was usually reinitialised at the beginning of each study list, as if the memory system were able to forget all previously learned items. This assumption predicts LLE and LSE because longer and stronger lists add variability to the decision process. Murdock and Kahana (1993a, 1993b) suggested that a more plausible assumption is *not* to reinitialise the composite vector at beginning of each list. The vector should also contain traces encoded prior to the experiment, as people have previous experience with the items being presented in the laboratory.

This *continuous memory* assumption readily explains the null LSE: the variance added by a few strong items during an experiment is simply not large enough to allow the detection of any difference between pure and mixed lists. In other

words, because there is so much variability from previous memories accumulated during a lifetime, the increase in variance during an experiment is negligible. Hence, no LSE is expected. The same argument applies to length manipulations and no LLE is expected either. But TODAM can still predict an LLE through its forgetting parameter  $\alpha$  ( $< 1$ ): the higher the number of items between an item's study and its subsequent test, the lower the discriminability at test. However, if the number of intervening items between study and test is the same across lists of different sizes, then no LLE should be observed. Put another way, an LLE could only occur if long lists have a larger number of intervening items between study and test, on average, than short lists. This prediction, however, was disconfirmed (Ohrt & Gronlund, 1999; see also 1.3.1). Thus, TODAM with the continuous memory assumption cannot account for both a positive LLE and a null LSE.

#### 1.4.3. REM (Shiffrin & Steyvers, 1997)

The many problems facing global matching models called for a change in approach. A model named Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997) was proposed that could tackle most of those thorny issues. REM borrowed several elements from SAM, MINERVA2 and TODAM. Here we concentrate on a simplified version of REM (for a detailed description, with examples, see Shiffrin & Steyvers, 1997).<sup>7</sup>

In REM, items in memory are represented as vectors of features whose values range from 0 to  $\infty$ . Each non-zero, feature value is independently drawn from a geometric distribution; the probability that a feature takes value  $v$  is given by  $P(v) = g(1 - g)^{v-1}$ , where  $0 < g \leq 1$  and  $v > 0$ . For a fixed  $g$ , low feature values are more likely than high values.<sup>8</sup> Variance is introduced in the model through a noisy encoding process. Each memory trace is initialised with all features set to 0. During study, an incomplete copy of the item is stored. For each feature, there is a probability  $u$  that a value is stored and a probability  $1 - u$  that the feature remains at 0. For each stored value, there is a probability  $c$  that it is the same as

<sup>7</sup> A model similar to REM, called Subjective Likelihood Model (SLiM; McClelland & Chappell, 1998), has been developed independently and at the same time. For the sake of brevity, we focus here on REM only (for a comparison between REM and SLiM, see Criss & McClelland, 2006).

<sup>8</sup> Low  $g$  simulates low-frequency words, as they are more likely to have rare, high feature values (probability distribution is almost uniform). High  $g$  simulates high-frequency words, as they are more likely to have common, low feature values (distribution becomes positively skewed).

the corresponding feature value in the study item and a probability  $1 - c$  that the value is chosen at random from the geometric distribution. Thus, there are two sources of noise during encoding, as a feature may be either not stored or stored with the wrong value. Once a feature value is stored, however, it does not change. As a result, strengthening an item causes more values to be stored (replacing the zeroes of the remaining features) but does not alter the values of previously stored features. The assumption that the same trace is updated with every additional presentation of an item contrasts with the assumption of previous models (e.g. SAM, MINERVA2), where additional study entailed the encoding of a new copy of the item.

The test item (vector) is matched in parallel to all traces stored during study. REM is thus a global matching model. Each feature of the test item is independently matched to the corresponding feature of a trace in memory. Trace feature values can match the features of the test item either because the trace in fact corresponds to the test item or simply by chance. Conversely, trace feature values can mismatch the features of the test item either because the trace does not correspond to the test item or because the trace does correspond to the test item but the value of that particular feature was wrongly stored at study.

It is possible to formalise this probabilistic matching process with the concept of likelihood (the probability of observing the data given a hypothesis). In this case, “data” are the feature values of the test item and “hypothesis” can be either “the test item was studied” or “the test item was not studied”. By taking the ratio of the likelihood that the item was studied to the likelihood that it was not, one can obtain an index of which hypothesis is more likely to be true for a given test item. This likelihood ratio (match) of a test item  $j$  to a stored trace  $i$  is given by:

$$\lambda_{ij} = (1 - c)^{nq_{ij}} \prod_{v=1}^{\infty} \left[ \frac{c + (1 - c)g(1 - g)^{v-1}}{g(1 - g)^{v-1}} \right]^{nm_{ij}(v)} \quad (1.5)$$

where  $nq_{ij}$  is the number of non-zero mismatches between trace  $i$  and test item  $j$  and  $nm_{ij}(v)$  is the number of non-zero matches with value  $v$  (features with a value of zero do not contribute to the matching process). Note that the higher the matching value  $v$ , the higher the overall match ( $\lambda$ ). This implements the idea that

less common feature values (i.e., high  $v$  values, present in low-frequency words, for example) are more diagnostic during the matching process than more common feature values (i.e., low  $v$  values, present in high-frequency words). To combine the matching information from each trace into a single index, one can take the average of the likelihood ratios corresponding to the  $N$  stored traces:<sup>9</sup>

$$\Lambda_j = \frac{1}{N} \sum_{i=1}^N \lambda_{ij} \quad (1.6)$$

This index corresponds to the *odds* that the test item is old versus new (see Shiffrin & Steyvers, 1997, for a derivation). If the odds are greater than a criterion (e.g., criterion = 1) then an “old” response is produced. Note that, unlike previous global matching models, where decisions were made over a familiarity scale, the decisions in REM are made over an odds scale.

REM incorporates the concept of differentiation, which was useful in accounting for the null LSE (see 1.4.1), because it treats every additional presentations of a study item as another opportunity to encode a feature not previously encoded. This means that stronger items have a more complete and accurate representation in memory (more non-zero features) than weak items. The presence of more non-zero features has two consequences: first, the match ( $\lambda$ ) between a test item and its *own* representation in memory is stronger, as there are more features contributing to the matching process and their values are likely to match; second, the match between a test item and any *other* item in memory is weaker, as there are more features contributing to the matching process and their values are likely to mismatch. For strong targets, matching dominates mismatching and the overall odds ( $\Lambda$ ) increase, resulting in an increase in hits. For distractors, mismatching dominates and the overall odds decrease, resulting in a decrease in false alarms.<sup>10</sup>

Differentiation can account for the null LSE because both hits and false alarms decrease in tandem with the strength of other list items without any change in

---

<sup>9</sup> In REM papers, odds are represented by the letter  $\Phi$ . In this thesis, we represent odds with the letter  $\Lambda$  and reserve  $\Phi$  to symbolise the cumulative normal distribution function (see Chapter 2).

<sup>10</sup> This pattern of higher hits and lower false-alarms in a strong condition compared to a weak condition is called the *strength-based mirror effect*. REM explains it through differentiation with no change in decision criterion. By contrast, some models (e.g., dual process: Cary & Reder, 2003; single process: Stretch & Wixted, 1998a) can only accommodate the strength-based mirror effect by assuming a criterion shift without explaining why the shift occurred in the first place.



overall discrimination ( $d'$ ). In a mixed list, some items are strengthened at study (strong items) and some are not (weak items). The decrease in hits for weak items in a *mixed* list compared to items in a *pure weak* list occurs because the average match of a weak target to the strong traces in memory shrinks (strong items are more distinct). The decrease in false alarms occurs for the same reason: the match between distractors and strong traces drops. Thus, no difference in discrimination is expected for weak items between pure and mixed lists. The same applies to the comparison of strong items between mixed and pure lists.

However, REM predicts a positive LLE. Hits decrease and false alarms increase in the long list condition. Hits decrease because more items on the list reduce the odds elicited by a *target* test item. False alarms increase because each new stored item raises the possibility of an accidental match with a *distractor* test item. In addition, REM predicts a constant new-to-old variance ratio across list lengths and strengths, which is consistent with previous findings (Ratcliff et al., 1992; Ratcliff et al., 1994). The reasons behind this constancy of ratios are less clear (see Shiffrin & Steyvers, 1997, p. 150-151, for a discussion).

To summarise, REM represented a new breed of global matching models capable of addressing some of the difficulties facing previous models. It incorporated differentiation (by assuming that repetitions update the same trace in memory) and a Bayesian decision process (by assuming that recognition is based on the odds that an item is old rather than on the strength of its familiarity signal).

## 1.5. Empirical evidence: recent studies

In the previous session, we briefly described a model designed to account for, among other things, the positive LLE and null LSE. To the dismay of memory theorists, however, two recent studies have cast fresh doubts on the status of list-length and list-strength effects in recognition. Dennis and Humphreys (2001) found neither an LLE nor an LSE when several confounding variables (e.g., study-test lag, attention level, rehearsal redistribution and context reinstatement) were controlled at the same time. In addition, Norman (2002) found a reliable LSE in item recognition when a particular combination of encoding time,

encoding task, strength level and lure type were used. Here we present those two studies in some detail, as they form the basis of the research reported in this thesis, and discuss some additional evidence that further challenges the assumptions behind global matching models.

### 1.5.1. Evidence against list-length effects

Dennis and Humphreys (2001) reviewed the literature on list-length effects in recognition and concluded that many results previously interpreted as evidence for an LLE were marred by confounds that could lead to artifactual effects. An artifactual LLE occurs when the LLE is caused by the confounding variable, not by the presence of additional items on the list.

The first confound pointed out by Dennis and Humphreys (2001) is study-test lag. This confound can be controlled by using either a proactive or a retroactive design (see 1.3.1). Studies using retroactive design tend to show smaller LLEs. For example, Murnane and Shiffrin (1991a) found highly significant LLEs in most of their experiments when using a proactive design but only a marginal LLE when using a retroactive design (Exp. 3).

The second confound is attention level: participants may pay less attention to items at the end of a list than to items at the beginning of the list. To the extent that this happens, it affects long lists more heavily than short lists, especially in proactive designs, where the items of interest are located at the end of the long list. Moreover, differences in attention between short and long lists should be more pronounced when there is no encoding task requiring participant's engagement during item presentation. Dennis and Humphreys (2001) argued that the LLE observed by Ohrt and Gronlund (1999, Exp. 1), may have fallen prey to such problem (even though Ohrt and Gronlund did control study-test lag).

The third confound pointed out by Dennis and Humphreys (2001) is rehearsal redistribution (for a discussion of redistribution in the context of list-strength manipulations, see 1.3.2). Rehearsal redistribution occurs when participants use the retention interval to rehearse previously studied items. This may happen in both retroactive and proactive designs. If it happens, it is more likely to benefit

short lists because there are a greater proportion of items in a short list prone to receive the additional rehearsal time than in a long list. This, coupled with the fact that only a fraction of the long list is tested (to avoid test length effects; Schulman, 1974) can harm performance in long lists. Dennis and Humphreys (2001) suggested that redistribution can be reduced by using a retroactive design, by testing only a fraction of the items studied earlier on the list and by adopting an engaging distractor task. Testing a fraction of early items should reduce the advantage of short lists, as some of the rehearsed items will not be tested. Adopting an engaging distractor task should discourage rehearsal, especially in the short condition where retention interval is longer, because it would presumably keep participants mentally busy.

The final confound pointed out by Dennis and Humphreys (2001) is contextual reinstatement. This refers to the theoretical notion that recognition judgements may involve not only the matching of the test item to the traces of studied items but also the matching of the *test context* to the *study context* encoded when the item was stored. Context is a broad concept that includes both internal states (e.g., body temperature, transitory thoughts, cognitive strategies) and external states (e.g., illumination in experimental room, colour of item's font, properties of adjacent items on a list). Context is also assumed to gradually change over time, as internal and external states change. Dennis and Humphreys (2001) argued that, when the retention interval is short (e.g., 10 s), an LLE can be generated by a form of context inertia. An LLE generated by such inertia would be artifactual as it would not be caused by interference from the other list items.

When retention interval is short, there is little time for the study context experienced by the participants to change. Consequently, participants may continue to use at test the same type of internal information they were using during study. This context inertia is beneficial to items studied late on the list, as test and study context are somewhat similar. Context inertia, however, is harmful to items studied early, as test and study contexts are dissimilar.

When retention interval is long (e.g., > 60 s), participants are obliged to *reinstate* the study context from the cues at hand (i.e., test item, test instructions), as the current encoding context is no longer useful. This extra processing effort may

benefit performance. Context reinstatement is beneficial to items studied early on the list, as they profit the most from a break in the test context, and it can also benefit late items if there is a long filled interval allowing context to change.

Context inertia (or lack of context reinstatement) can cause an artifactual LLE because it is more harmful to long lists than short lists, especially in studies using a retroactive design. Short lists are followed by a long interval; the context at test is therefore different, forcing participants to reinstate the original study context. Long lists, on the other hand, are followed by a short retention interval; the context at test is therefore similar to the context associated to items studied later on the list, harming recognition of early items. Dennis and Humphreys (2001) argued that such context inertia could explain the LLE observed by Gronlund and Elam (1994) because they used a retroactive design and a short retention interval (9 s in the long condition; 69 s in the short condition). Adopting a proactive design does not solve the problem. Although retention interval in the proactive design is the same for short and long lists (eliminating differences in context reinstatement), there is still the problem of attention loss (i.e., poor encoding) of late items following the study of a long list.

Following those considerations, Dennis and Humphreys (2001) undertook to test whether an LLE would still be observed after *all* confounds were controlled at the same time. They carried out two experiments neither of which showed any hint of an LLE; their Experiment 2 also showed no LSE. Both a null LLE and a null LSE are direct predictions from their model called BCDMEM. We describe Dennis and Humphreys' (2001, Exp. 2) findings here and their model in 1.6.1.

Dennis and Humphreys (2001, Exp. 2) carried out a study where short lists contained 40 items and long lists contained 80 items. Moreover, the experiment included a mixed-strength list containing 10 items presented once and 30 items presented three times. Study-test lag was controlled with a retroactive design. Attention loss was controlled, since only early items were tested in both lists. Rehearsal redistribution was controlled because they adopted an engaging puzzle task as filler and tested only a fraction of the items studied in the first half of the study phase. Finally, contextual reinstatement was controlled because they

provided a long retention interval for both short and long lists (360 s and 240 s, respectively). Apart from the expected criterion shift in the mixed-strength list (fewer hits and false alarms), there was no difference between the conditions.

The conclusions about the lack of LLE and LSE rely on the acceptance of statistical null hypotheses, which raises the issue of statistical power. According to Frick (1995), a null hypothesis can be confirmed when a reasonable amount of effort is put into finding the corresponding effect. Additional support for a null effect may be obtained if another manipulation, run concurrently with the manipulations of interest, is shown to produce an effect. Dennis and Humphreys (2001) performed an orthogonal word frequency manipulation to show that their design had sufficient power and indeed found a significant effect (i.e., low-frequency words were better recognised than high frequency words, a phenomenon known as the *word-frequency mirror effect*). The significance level associated with the effect, however, was modest ( $p = .03$ ) considering the usual size and reliability of this effect (Glanzer & Adams, 1985, 1990). In addition, the length manipulation used by Dennis and Humphreys (2001) was perhaps not as strong as in several other studies. The length ratio between long and short list was 3:1 (Exp. 1) and 2:1 (Exp. 2), compared to the ratios used by Ohrt and Gronlund (1999; 5:1 in Exp. 1 and 4:1 in Exp. 2). In short, it is possible that the null results were obtained due to a lack of power.

Indeed, Cary and Reder (2003, Exp. 3) replicated Dennis and Humphreys' (2001) design and found an LLE. Their main goal was to show that list-length effects are a real phenomenon in recognition and that it can be modelled using a dual-process memory model (see description of SAC in 1.6.2). Two design features adopted by Cary and Reder (2003, Exp. 3) and not by Dennis and Humphreys (2001) could account for the conflicting results: the use of the *Remember/Know* paradigm and the use of a 4:1 length ratio.

In the *Remember/Know* (RK) paradigm (Tulving, 1985), participants are asked to report their state of awareness associated with each “old” response they make. *Remember* responses are assumed to reflect the recall of specific details about the encoding event; *Know* responses are assumed to reflect the feeling of familiarity

associated with the event in the absence of recall. Cary and Reder's (2003) results may reflect a disproportional use of recall during the recognition task as the RK instructions may cause participants to rely more heavily on recall than they would normally do. The disproportional presence of recall in recognition could lead to an LLE because the effect is found in recall and cued recall.

To test the possibility that Cary and Reder's (2003) use of the RK procedure was the reason behind their positive LLE, Kinnell and Dennis (2007) carried out an experiment comparing performance between the standard old-new recognition test and a recognition test using the RK paradigm. In both cases, no LLE was found. The result suggests that the RK procedure is not the critical factor behind Dennis and Humphreys' (2001) and Cary and Reder's (2003) discrepant results. The remaining possibility, the different length ratios adopted in those studies, is one of the variables manipulated in the research reported in this thesis.

To summarise, Dennis and Humphreys (2001) identified a series of possible confounds that could have accounted for previous list-length effects. They carried out two carefully designed experiments aimed at removing those confounds and found no evidence of LLE and LSE. A similar study later conducted by Cary and Reder (2003), however, did find an LLE. Taken together, these results suggest that, although an LLE can be found in recognition, the boundary conditions under which it is produced are still not clear.

### 1.5.2. Evidence for list-strength effects

The failure to find an LSE in recognition was taken as evidence that the effect does not exist. New global matching models, such as REM (Shiffrin & Steyvers, 1997), were developed to account naturally for this null result. In this context, the finding by Norman (2002) that an LSE could in fact be produced in recognition may have come as bad news for some theorists.

Norman (2002) reasoned that, because LSE is observed in free recall and cued recall, it should also be found in recognition if participants are forced to use a recall-like process during the recognition decision. This dual-process view of recognition memory assumes that an "old" response may be elicited either by a

general feeling of familiarity or by a process akin to recall called *recollection*. In the following, we describe Norman's (2002, Exp. 2) design. Evidence supporting dual-process models of recognition are reviewed in the next section.

Norman (2002, Exp. 2) forced the use of recollection at test by using test lures that were very similar to studied targets. He adopted the switched-plurality paradigm (Hintzman, Curran, & Oppy, 1992), where participants are presented with items at study in a given plurality (e.g., singular: *banana*; plural: *trees*) and are tested with the same items either in the same plurality or with their plurality switched (e.g., *bananas*, *tree*). Participants are instructed to pay attention during study whether the item is singular or plural. They are also informed that they may see either the same item at test or the item with its plurality reversed. The task forces participants to rely on a sense of recollection because a simple feeling of familiarity does not allow discrimination between targets and lures at test: *banana* and *bananas* have similar levels of familiarity. So participants can only choose one version over the other by remembering which version was studied.

Norman (2002) argued that previous experiments did not produce an LSE because participants were not required to use recollection at test. Moreover, he argued that the strength manipulation of previous studies was too weak and that the encoding times and encoding tasks used in those studies were not optimal. He then took measures to increase the chances of observing the effect. In his study, the number of study presentations was high (5 repetitions), the encoding time was short (1.15 s) and the encoding task was demanding (size judgement).

Increasing the number of presentations should bolster interference according to some models (see 1.6.2 and 1.6.3; though REM predicts no change). To allow the increase in the number of repetitions, Norman (2002) focused on the comparison of weak items between *pure weak* lists and *mixed* lists. The complementary comparison (strong items from *pure strong* and *mixed* lists) would not be informative as performance for strong items would be at ceiling in both lists.

Setting the encoding time at the right level is also relevant because this is a sensitive parameter: if encoding time is too short (e.g., < 1 s), the stored trace

may be too impoverished to allow recollection to operate at test (Gardiner & Gregg, 1997); if it is too long (e.g.,  $> 3$  s), the trace may become too differentiated to interfere with other list items (e.g., Ratcliff et al., 1990, Exp. 7). In both cases, a positive LSE would be more difficult to find. In his doctoral dissertation, Norman (1999) reported data obtained under different encoding times and found that 1.15 s produced the best results.

Finally, choosing the correct encoding task may increase the chances of producing interference between memory traces. The encoding task used by Norman (2002) required participants to judge, for each study item, whether or not a typical instance of that item would fit into a banker's box present in the experimental room. As argued by Norman (2002), the purpose of the task was threefold: *i*) to allow participants to encode the items deeply, increasing the chances of recollecting them at test; *ii*) to force participants to pay attention to the words during encoding, reducing attention loss; *iii*) to increase the chances of memory interference, as all words would be encoded with respect to the same referent (i.e., the banker's box). In contrast, most previous studies used either no encoding task at all (participants were just told to memorise the items) or less optimal encoding tasks (e.g., pleasantness rating).

At test, participants in Norman (2002, Exp. 2) were presented with three types of items: targets (e.g. *banana*), switched-plurality lures (e.g. *bananas*) and unrelated lures (e.g. *car*). According to the dual-process hypothesis, recollection is affected by strength interference whereas familiarity is not. If this hypothesis is correct, then an LSE should be observed with switched-plurality lures, where recollection is more likely to operate, but not with dissimilar lures, where recollection is less likely. In two experiments, Norman (2002) observed exactly this pattern.

Diana and Reder (2005) used the RK paradigm and replicated some of Norman's (2002, Exp. 1) results. They found a decrease in *Remember* responses of weak items from *mixed* lists compared to weak items from *pure weak* lists. However, they did not find a significant decrease in  $d'$  for those *Remember* responses (i.e., when *Remember* false alarms are taken into account, the previously observed



LSE goes away).<sup>11</sup> The same results were found in a second experiment in which strong items were presented 11 times. Diana and Reder (2005) argued that the failure to replicate Norman (2002, Exp. 1) was probably caused by differences in the experimental designs used in those studies.

Like Norman (2002, Exp. 1), Diana and Reder (2005) used the RK paradigm to obtain separate estimates of recollection and familiarity. Also, only unrelated lures were used at test. Unlike Norman (2002, Exp. 1), however, they presented items for longer times (1.5 s) and used an arguably less powerful encoding manipulation (participants were asked to decide whether the font of the item was appropriate for the item's meaning). More importantly, Diana and Reder (2005) adopted a long retention interval (5 min. in the *mixed* lists); long intervals may reduce the possible interfering effects of strong items. The role of retention interval in LSE is another variable investigated in the research presented here.

In sum, Norman (2002) carried out two experiments carefully designed to maximise the chances of observing an LSE. The rationale behind those experiments came from dual-process theories, according to which two different processes, familiarity and recollection, underlie recognition judgements. Because recollection is akin to recall and because LSE is observed in recall, LSE should be found in recognition if recall plays a sufficiently relevant role during test. Diana and Reder (2005) partially replicated Norman's (2002) results but differences in the experimental designs prevent one from drawing strong conclusions. Together, these results suggest that an LSE can be found in recognition but it is still unclear which conditions are essential for its production.

### 1.5.3. Recent challenges to global matching models: the role of recall

The findings of a null LLE (Dennis & Humphreys, 2001) and a positive LSE (Norman, 2002) cast some doubt over the empirical status of those effects in

---

<sup>11</sup> The null Remember  $d'$  was caused by a simultaneous decrease in Remember hits and false alarms. This is consistent with the view that Remember/Know responses simply reflect the placement of different criteria along the same familiarity dimension rather than reflecting the workings of two qualitatively different memory processes (e.g., Donaldson, 1996; Dunn, 2004). According to this view, hits and false alarms changed together in Diana and Reder (2005) because the Remember criterion shifted across conditions. Yet this interpretation is clouded by the fact that no change in Know responses was found between pure and mixed list conditions.

recognition. In addition, further findings suggesting that the new-to-old variance ratio is *not* constant across length and strength manipulations (Heathcote, 2003) called into question previous results showing constant variance ratios (Gronlund & Elam, 1994; Ratcliff et al., 1992; Ratcliff et al., 1994). These results are theoretically relevant because modern global matching models, such as REM, were designed to account for the null LSE and the constancy of variance ratios.

More challenging to global matching models, however, is the growing body of evidence converging on the view that a recall-like mechanism operates during recognition. Evidence consistent with this view comes from behavioural (e.g., Boldini, Russo, & Avons, 2004; Hintzman & Curran, 1994; Rotello, Macmillan, & Van Tassel, 2000), physiological (e.g., Curran, 2000; Rugg & Curran, 2007), neuropsychological (e.g., Bowles et al., 2007; Mayes et al., 2002) and pharmacological studies (e.g., Curran, DeBuse, Worch, & Hirshman, 2006; Hirshman et al., 2002). We review here some of that evidence and discuss why a recall mechanism may be relevant to LLE and LSE in recognition.

### Familiarity and recollection

The view that two processes underlie recognition memory is not new (Atkinson & Juola, 1974; Mandler, 1980; see Yonelinas, 2002, for a review). However, only recently sufficient behavioural and neurophysiological evidence became available to provide a viable alternative to the single-process view. Dual-process models assume that two qualitatively different sources of memory information, namely, familiarity and recollection, determine performance during a recognition test. Although there are many different dual-process models, each with its own definitions and implementations of familiarity and recollection processes (Some-or-None: Kelley & Wixted, 2001; CLS: Norman & O'Reilly, 2003; SAC: Reder et al., 2000; STREAK: Rotello, Macmillan, & Reeder, 2004; All-or-None: Yonelinas, 1994), they generally agree on the basic properties of each process.

*Familiarity* is conceptualised as a fast, context-insensitive, automatic process, whereas *recollection* is a slow, context-sensitive, strategic process. Moreover, the two processes seem to operate in a fall-back manner, whereby decisions are

based on familiarity only if recollection fails. Decisions may be based solely on familiarity if recollection is impoverished (e.g., long study-test lag), if familiarity alone is discriminative (e.g., target vs. new pair discrimination) or if there is time pressure (e.g., short lags in response-time studies). Many recognition memory results can be accounted for by assuming that familiarity and recollection differentially contribute to performance in a given task. In the following, we briefly review some evidence from associative recognition and item recognition studies that support the existence of recollection.

### Associative recognition

The first indication that recall could contribute to recognition came from *associative recognition* studies. In this paradigm, participants study pairs of items (e.g., AB, CD) and are tested with previously studied pairs (e.g., AB; *targets*), new pairs made up of previously studied items (e.g., AC; *rearranged pairs*) or new pairs made up of previously unstudied items (e.g., EF; *new pairs*). Early results indicated that associative recognition resembled recall more than recognition. For instance, the memory advantage for low-frequency words found in item recognition is reversed in associative recognition as it is in recall (Clark, 1992). Also, the time course of associative recognition is more similar to the time course of cued recall than that of item recognition (Nobel & Shiffrin, 2001).

The most suggestive early evidence that a recall-like process could be operating in associative recognition came from *response-signal experiments* (Reed, 1973). In these experiments, participants are given a signal to respond “old” or “new” on each trial and have to respond immediately after the signal. The lag between the onset of the test item and the onset of the signal varies from trial to trial: sometimes the lag is short (e.g., 100 ms), sometimes it is long (e.g., 1000 ms). The technique allows the identification of the point at which a recall-like mechanism becomes available at test, if it indeed contributes to performance.

Gronlund and Ratcliff (1989) used the response-signal technique to investigate discrimination between target pairs (e.g. AB, CD), rearranged pairs (e.g., AC) and new pairs (e.g., XY). They found that false alarms to new pairs first

increased with increasing lags and then started to decrease at about 350 ms, whereas false alarms to rearranged pairs continued to increase until about 600 ms. These results can be interpreted as evidence for the operation of a recollection mechanism in associative recognition. The first false-alarm peak indexes the familiarity process. Because neither item in the new pair has been previously studied, familiarity alone is discriminative: if the signal from each item does not reach criterion, a “new” response is given. Recollection of associative information is therefore not necessary. The second false-alarm peak indexes the recollection process. Recollection allows discrimination between target and rearranged pairs, as it enables the participant to reject the rearranged pairs: the participants can use A to recall B and thus classify the test pair AC as “new”, a process called *recall-to-reject*. Familiarity alone is not sufficient to allow discrimination because both items in the test pair have been previously studied and, consequently, should elicit similar levels of familiarity.

Rotello and Heit (2000) provided converging evidence that recall-to-reject operates in associative recognition by using SDT measures that take into account changes in response criterion. At short lags, participants have little information to base their decisions on; as a consequence they may set a lenient criterion to respond “old”. As more information becomes available at longer lags, participants may raise their response criterion (see Heit, Brockdorff, & Lamberts, 2003, for evidence of change in criterion setting across response-signal lags). This criterion shift, from liberal to conservative, could account for the decreases in false alarms across lags. Rotello and Heit (2000) argued that a more specific measure of recall-to-reject involves directly comparing false-alarm rates from rearranged and new pairs. Both should move together with criterion shifts but only rearranged pairs are sensitive to recall-to-reject. The difference in the proportion of false alarms should change when recall-to-reject begins to operate and this difference can be measured with a sensitivity index akin to  $d'$ . Using this measure, Rotello and Heit (2000) found reliable recall-to-reject in associative recognition. Similar results were obtained in a subsequent study using ROC curves (see Chapter 2), which also takes criterion shifts into account: when response-signal lag was long (2500 ms), recall-to-reject was observed; when lag was short (450 ms), recall-to-reject disappeared (Rotello et al., 2000, Exp. 4).

Because list-length and list-strength effects are found in recall and because associative recognition appears to have a recall component, it stands to reason that LLE and LSE should also be found in associative recognition if the appropriate conditions for recollection are provided (i.e., enough time to respond and enough target-lure similarity to allow the use of recall-to-reject). Consistent with this prediction, both LLE (Criss & Shiffrin, 2004c) and LSE (Verde & Rotello, 2004) have been found in associative recognition.

### Item recognition

The evidence supporting the role of recollection in item recognition has a more tortuous history. Hintzman and Curran (1994) used the response-signal technique to assess the temporal dynamics of recognition and found a result similar to Gronlund and Shiffrin's (1989): false alarms peaked earlier for unrelated lures than for switched-plurality lures. The late peak was taken as evidence for a slow, recall-to-reject process. However, when those results were reanalysed by comparing the fits of a monotonic model, which suggests no recall-to-reject, against a non-monotonic model, which suggests the presence of recall-to-reject, the results favoured the former (Rotello & Heit, 1999). This contrasted with evidence for recall-to-reject in associative recognition (Rotello & Heit, 2000). One possible reason for the discrepancy is that recall-to-reject may be differently recruited in item and associative recognition. Westerman (2001, Exp. 3) showed that priming subsequently tested items increased false alarms to switched-plurality lures in item recognition but not to rearranged lures in associative recognition. This suggests that the increased familiarity from priming was less effectively counteracted by recall-to-reject in the former than in the latter.

Although associative and item recognition do behave differently in many instances, it is possible that recall-to-reject in item recognition is under strategic control and, consequently, would require more direct instructions in order to be elicited. Participants in Westerman (2001) were not explicitly told that recalling an item in one plurality meant that the alternative plurality was not studied (although they were told to pay attention to the plurality of the word at study and

to say “old” only if word and plurality matched). Participants in Hintzman and Curran (1994, Exps. 2 and 3) were also not explicitly told to use a recall strategy. Rotello et al. (2000) tested the hypothesis that recall-to-reject in item recognition is under strategic control by either telling participants to use a recall strategy (Exp. 1) or not (Exp. 2). In the first case, participants were told that if *banana* is recalled, then *bananas* could not have been studied, and a *definitely new* response should be given. In the second case, participants were only told to say “old” to studied words and reject all others. The results showed larger estimates of recall-to-reject in Exp. 1 than in Exp. 2, consistent with the hypothesis that recall-to-reject in item recognition is modulated by strategic processes. Taken together, these results suggest that recollection, in the form of a recall-to-reject strategy, also occurs in item recognition (as it does in associative recognition).

Another line of evidence supporting the dual-process view comes from studies in which variables thought to differentially affect familiarity and recollection are manipulated. Familiarity is thought to be more sensitive than recollection to perceptual features of the study item (Toth, 1996) and recollection is thought to be more sensitive than familiarity to its semantic features (Gardiner, 1988). Boldini, Russo and Avons (2004) manipulated perceptual features by varying study-test modality (auditory-visual vs. visual-visual) and semantic features by varying depth of processing (shallow vs. deep). The results showed that discriminability  $d'$  was higher when study-test modality matched than when it mismatched, but only at short response lags ( $\leq 300$  ms). By contrast,  $d'$  was higher when encoding was deep compared to shallow, but only at long response lags ( $> 300$  ms). The dissociation is consistent with the idea that recognition can be based on two processes – one fast-acting and one slow-acting – which can be differently affected by different properties of the stored traces.

Recall-to-reject and response-time dissociations are not the only measures of familiarity and recollection in recognition. Other methods, such as the process-dissociation procedure (Jacoby, 1991; Jacoby, Toth, & Yonelinas, 1993), the ROC procedure (Yonelinas, 1994, 1997) and the *Remember/Know* procedure (Gardiner & Richardson-Klavehn, 2000; Tulving, 1985) have also been used. These methods, however, are based on assumptions that sometimes are not met,

and, therefore, received a considerable amount of criticism (Curran & Hintzman, 1995; Donaldson, 1996; Dunn, 2004; Heathcote, 2003; Wixted, 2007).

Nevertheless, the estimates obtained from these procedures all converge on the view that recognition is mediated by two processes.<sup>12</sup>

Estimates from the ROC and RK procedures were used to assess the differential impacts of list-length and list-strength manipulations on familiarity and recollection. Yonelinas (1994) found an LLE for the recollection component of his model but not the familiarity component. Likewise, Norman (2002) found an LSE for *Remember* responses but not for *Know* responses. Both results are consistent with the idea that the LLE and LSE found in free recall, cued recall, associative recognition and item recognition are mediated by a similar mechanism that is sensitive to interference from the other traces in memory.

These process estimation methods have also been used in the study of the neurobiological correlates of familiarity and recollection (Rugg & Yonelinas, 2003). Event-Related Potentials (ERP) provided evidence consistent with the conclusions from response-signal studies. ERPs have good temporal resolution, thereby allowing the study of the time-course of recognition. In an item recognition experiment using switched-plurality lures, Curran (2000) showed that an early ERP component (FN400; 300-500 ms) was modulated by the familiarity of the test item (new items elicited a lower signal than studied and switched-plurality items), whereas a late ERP component (parietal; 400-800 ms) was modulated by the recollection of the item's plurality (studied items elicited a higher signal than switched-plurality and new items). The differences in timing and topography of the ERP components suggested that not only two processes could be operating in item recognition but also that they may be mediated by different underlying neural systems.

Norman et al. (2008) found ERP evidence that a list-strength manipulation selectively affects recollection but not recognition. The late parietal component (the hypothesised index of recollection) elicited by weak items was smaller in the

---

<sup>12</sup> See Yonelinas (2002) for a thorough review of manipulations that differentially affect estimates of familiarity and recollection derived from process-dissociation, ROC and RK.

*mixed* list than in the *pure weak* list, whereas the FN400 component (the hypothesised index of familiarity) did not change across lists. The result was replicated in a second experiment using the RK procedure: both *Remember* responses and the late parietal component decreased for weak items in *mixed* lists; both *Know* responses and the FN400 component remained unchanged.

In sum, behavioural and neurobiological evidence gathered over the last 20 years have strengthened the case that two processes may operate in item recognition. In particular, the evidence is consistent with the view that list-length and list-strength manipulations affect recollection more than familiarity.

### Single-process accounts

Despite the wealth of evidence supporting a dual-process view of recognition, a consensus has yet not been reached partly because single-process models are considered more parsimonious and are still able to account for many extant data. Below, we describe a few examples of results that were taken as evidence for dual-process models but that can also be handled by single-process models.

The evidence from response-signal experiments (Boldini et al., 2004; Hintzman & Curran, 1994) can be accounted for by a dynamic, single-process, recognition model that assumes that participants base their decisions on information that changes over time (FESTHER; Brockdorff & Lamberts, 2000; Lamberts, 1995). The similarity (match) between the test item and the stored traces may change over time as more and more features become available. Perceptually salient features are available earlier and command greater weight on early decisions. Words sharing similar perceptual features, such as *banana/ bananas*, or words presented in the same modality at study and test, tend to have greater similarity early on. Greater similarity implies higher familiarity and higher probability of responding “old”. This could account for the initial increase in switched-plurality false alarms in Hintzman and Curran (1994) and for the initial increase in ‘visual-visual’ hits in Boldini et al.’s (2004) study-test modality manipulation. Later on during the recognition event, when less salient features become more available, the perceptual similarity between test item and memory traces may



decrease, whereas their semantic similarity may increase. This could account for the late decrease in switched-plurality false alarms in Hintzman and Curran (1994) and for the late increase in ‘deep’ hits in Boldini et al.’s (2004) levels-of-processing manipulation. Thus, data from response-time studies, normally taken to support dual-process models, can be explained by a single-process model.

Another example of a single-process account comes from studies on the word-frequency effect. Previous research has shown that manipulations thought to affect recollection more than familiarity, such as study-test lag or encoding task, affected the low-frequency hit rate advantage but not the low-frequency false-alarm advantage (Joordens & Hockley, 2000). The results indicated that the hit and false-alarm portions of the word frequency effect may be mediated by different processes. In particular, the results suggested that recollection underlies the hit-rate portion of the effect. Further support for this claim came from a study in which participants performed a recognition test after taking either saline or Midazolam, a benzodiazepine drug that causes anterograde amnesia (Hirshman et al., 2002). The assumption was that Midazolam is particularly harmful to recollection, since amnesic patients show higher impairment on estimates of recollection than familiarity (Yonelinas et al., 1998). Hirshman et al. (2002) showed that the low-frequency hit-rate advantage was reversed in participants under Midazolam, whereas the low-frequency false-alarm advantage remained unchanged. The result was taken as further support for the dual-process account of the word frequency effect. However, a subsequent study showed that the effect of Midazolam on the word frequency effect could be modelled by the REM model simply by assuming that Midazolam decreases the probability  $c$  of accurately storing a feature at study (Malmberg, Zeelenberg, & Shiffrin, 2004). In other words, a recollection-free model can account for Hirshman et al.’s (2002) data by assuming noisier encoding of study word features.

Some long-time supporters of the single-process account, however, have recently conceded that, at least in some special circumstances (e.g., when lures are highly similar), a recall-like process is required to account for experimental data. A case in point is REM. One issue with REM, originally conceived as a single-process model, is that its matching process causes the likelihood ratio  $\lambda$  (and

consequently, the odds  $\Lambda$ ) to increase excessively with lure similarity (Criss & McClelland, 2006, Fig. 2). As a result, REM overpredicts false alarms in experiments where very similar lures are used, as the  $\lambda$  generated by the lure's partial match is too high. Malmberg, Holden and Shiffrin (2004) faced this problem when trying to fit the REM model to data replicating the *registration-without-learning* phenomenon (Hintzman et al., 1992), where discrimination between targets and switched-plurality lures does not improve with additional study, despite an increase in judgements of target frequency. REM overpredicted the false-alarm rate observed in the data because its matching process gave disproportional weight to the partial match from the highly similar lures. The model was able to fit the data, however, when a search process (akin to recall) was added, counteracting the increase in false alarms (i.e., the model was able to *recall* a target when presented with a similar lure, allowing it to reject the lure).

To summarise, single-process models of recognition are still able to fit many empirical results. However, there are situations where a single-process model may not be able to capture the regularities in the data unless it assumes a recall-like mechanism. Moreover, evidence from behavioural, pharmacological, neuropsychological and neurophysiological studies support the view that recognition may be implemented in the brain by more than a single familiarity signal. In the next section, we discuss some alternative approaches.

## 1.6. Alternatives: context-noise and dual-process models

The difficulties faced by the single-process models motivated the development of new models that either assumed a dual-process view from the outset (e.g., SAC and CLS) or assumed a single-process view but radically changed the way the familiarity signal is produced (e.g., BCDMEM). Below, we give a short description of those models, focusing on how they explain LLE and LSE.

### 1.6.1. BCDMEM (Dennis & Humphreys, 2001)

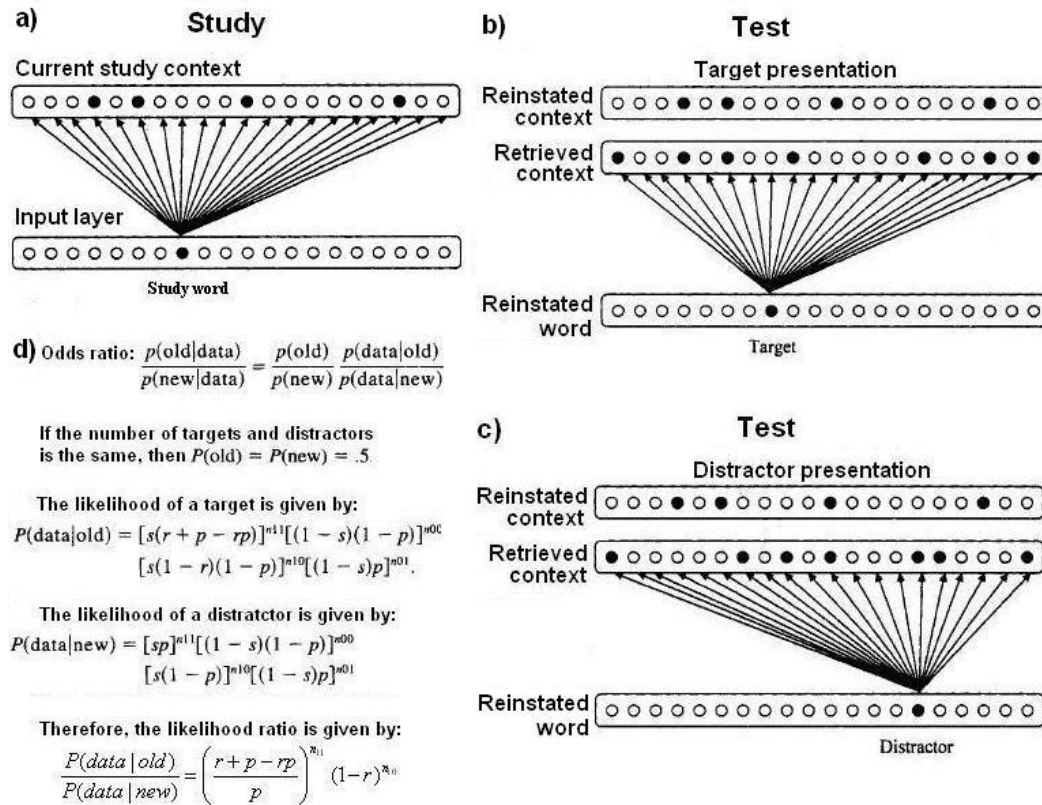
The Bind Cue Decide Model of Episodic Memory (BCDMEM; Dennis & Humphreys, 2001) assumes that recognition memory is a process sensitive to context interference and insensitive to item interference. *Context interference* is the interference (noise) generated by the different contexts in which an item has

been studied (e.g., prior to an item's presentation in an experiment, the item may have been seen in a magazine, in a book, on the TV, etc.). *Item interference*, on the other hand, is the interference generated by the other items present in the current context (i.e., the presence of many items or stronger items on a list adds noise to recognition). Item interference is assumed to be more relevant in recall.

Dennis and Humphreys' (2001) assumption that recall and recognition are susceptible to different sources of interference entails two predictions. First, list-length and list-strength effects should be observed in recall, insofar as the noise contributed by extra items or stronger items on a list should impair the memory of the remaining items in that list. Second, list-length and list-strength effects should *not* be observed in recognition, insofar as variations in list length and strength do not increase the context noise of the remaining items on the list. The prediction of no LLE and LSE in recognition is controversial. Nevertheless, Dennis and Humphreys (2001) found some evidence for their claim (see 1.5.1).

Figure 1.2 illustrates the BCDMEM model. Words are represented as individual nodes in the input layer and contexts are represented as a set of active nodes in the context layer. A node in the context layer can be either active (activation = 1; solid circles in Figure 1.2) or inactive (activation = 0; open circles). A node in the context layer has a probability  $s$  (sparsity) of being active at study. Each node in the input layer is connected (*bound*) to each node in the context layer through a set of associative weights. At study, the associative weight connecting an item to a context node is either set to 1 with probability  $r$  (learning rate) or kept at 0 with probability  $1 - r$  whenever the nodes are active at the same time. Thus not all contextual features active during study are encoded with the study item.

Recognition is instantiated in BCDMEM by presenting (*cueing*) the model with both the test item and the test context. The presentation of the test item activates the corresponding item node in the input layer, which in turn activates the context layer. The pattern of activation in the context layer is a composite containing the contexts with which the item has been previously associated. The probability  $p$  that a node is active due to previous learning is called context noise. The *retrieved context* is then compared to the test context (*reinstated context*).



**Figure 1.2. Bind Cue Decide Model of Episodic Memory (BCDMEM).**

(a) Each word is bound together with the context present at study. Encoding is probabilistic so that not all active context features are encoded with the word. (b, c) Recognition occurs by cueing the model with the test item (*reinstated word*) and the test context (*reinstated context*). The reinstated word prompts the model to retrieve its context vector, which is a composite containing all contexts where the word has been previously studied (*retrieved context*). The retrieved context is then compared with the reinstated context. The degree of match between the two vectors, which varies depending on whether the test item is a target (b) or a distractor (c), determines the recognition decision. (d) The decision mechanism is based on the odds that the item is old versus new. The odds vary as a function of the number of matches and mismatches between the vectors ( $n_{11}$ ,  $n_{10}$ ,  $n_{01}$ ,  $n_{00}$ ) and as a function of four parameters:  $s$  (vector sparsity),  $r$  (learning rate),  $p$  (context noise) and  $d$  (context reinstatement). If the odds are greater than 1, an “old” response is produced.

Adapted from Dennis and Humphreys (2001). © American Psychological Association.

To decide whether a test item was seen on the study list, the model has to take into account the probabilistic nature of the encoding and retrieval processes and the fact that the studied item has been encountered in different contexts prior to the experiment. Errors may be produced either because a distractor activates a context that shares many features with the current test context (spurious match; false alarm) or because a previously studied item has not encoded a certain feature of the study context (spurious mismatch; miss). There are four types of matches, namely, 11 (both reinstated and retrieved nodes are active), 10, 01, 00. Decisions are made in BCDMEM through a Bayesian mechanism similar to the

one adopted in REM (see 1.4.3). An odds ratio is calculated, reflecting the relative probabilities that an item is a target or a distractor given the number of matches and mismatches between the reinstated and retrieved contexts and given the learning rate and context noise. When there is no criterion manipulation (criterion = 1), the odds ratio reduces to a likelihood ratio and is given by:

$$\frac{P(data | old)}{P(data | new)} = (1-r)^{n_{10}} \left[ \frac{(r+p-rp)}{p} \right]^{n_{11}} \quad (1.7)$$

where  $n_{ij}$  are the number of  $ij$  matches between reinstated and retrieved contexts. Discriminability suffers when the learning rate ( $r$ ) approaches 0, indicating weaker encoding or when the context noise ( $p$ ) approaches 1, representing higher word frequencies. In both cases, the likelihood ratio approaches 1, reducing the model's ability to discriminate between targets and distractors.

Equation 1.7 was derived assuming that the reinstated context is an exact replica of the context present when the item was studied. It is likely, however, that the reinstated context is somewhat different due to factors such as delay. Dennis and Humphreys (2001) argued that as time passes, new context features may become active and old context features may become inactive, causing a reduction in the similarity between study and reinstated contexts. To take this contextual drift into account, Dennis and Humphreys (2001) introduced a context reinstatement parameter ( $d$ ) representing the probability that a node will change from active in the study context to inactive in the reinstated context. When  $d$  is introduced into the model, the likelihood ratio becomes:

$$\frac{P(data | old)}{P(data | new)} = \left[ \frac{1-s+d(1-r)s}{1-s+ds} \right]^{n_{00}} (1-r)^{n_{10}} \left[ \frac{p(1-s)+d(r+p-rp)s}{p(1-s)+dps} \right]^{n_{01}} \left[ \frac{(r+p-rp)}{p} \right]^{n_{11}} \quad (1.8)$$

Equation 1.8 reduces to 1.7 when  $d = 0$ . Note that now 01 and 00 matches are relevant to the likelihood ratio calculation because an inactive node in the reinstated context (activity = 0) may have been active at study (activity = 1).

BCDMEM is not a global matching model. The model differs from most previous models not only because it uses odds as the decision variable but also because the odds signal is based on the match between the retrieved and reinstated contexts of the test item, not on the match between all stored traces

and the test item. This occurs because BCDMEM activates the representation of the test item directly in memory, since words in the model are individual nodes.

BCDMEM can account for previous findings, such as the word frequency mirror effect and the list-length effect, in terms of its context noise and context reinstatement parameters, respectively. The model can account for the word frequency effect by assuming that context noise ( $p$ ) is higher for high-frequency than for low-frequency words, reflecting the fact that the former are more likely than the latter to be seen in several contexts (Steyvers & Malmberg, 2003). Dennis and Humphreys (2001) fitted the word-frequency data from Glanzer and Adams (1990) by fixing  $r$  and  $d$  and allowing  $p$  to vary. The estimated value of  $p$  was lower for the low-frequency words compared to the high-frequency words.

The model was also able to fit the list-length effect reported by Gronlund and Elam (1994). The model was fitted by having  $r$  and  $p$  fixed and allowing  $d$  to vary between short and long lists (Dennis & Humphreys, 1998). The best fit was obtained with higher  $d$  values for long lists compared to short lists, indicating that the reinstated context for long lists was a poor match to the retrieved context.

BCDMEM's fit to list-length data suggests that context noise, rather than item noise, is behind the list-length effect. Poor context reinstatement should occur in experiments using a retroactive design when retention interval is short. If a test is taken immediately after study, participants might continue to use the temporal context experienced at the end of the list, which is associated with items that are not tested, and fail to reinstate the context present at the beginning of list, which is associated with both tested and non-tested items. The longer the list, the larger the mismatch between beginning- and end-of-list contexts, and the larger this mismatch, the poorer the memory (Dennis & Humphreys, 2001, p. 458). Thus, although forgetting seems to be caused by interference from other items on the list, according to BCDMEM, it is a consequence of contextual drift over time.

BCDMEM is not the only model incorporating the idea that context noise is an important source of interference in recognition. Criss and Shiffrin (2004a) showed that a modified version of REM in which memory vectors contain

context features as well as item features can explain joint context and item noise effects. In that study, participants studied three lists of categorised items. Targets were defined as List 3 items; that is, participants were told to say “old” to List 3 items and “new” to all other items. Context noise was manipulated by repeating some study items on more than one list. Item noise was manipulated by increasing category length in the final list. Context interference occurred because false-alarm rates increased with the number of previous contexts in which a word has been studied (i.e., false alarm was higher for items studied in Lists 1 and 2 than for items studied in either List 1 or 2). Item interference occurred because false-alarm rates increased with category length. The results suggest that both item and context noise are important sources of interference in recognition.

In summary, BCDMEM was developed as an alternative to global matching models that emphasises the importance of context interference in recognition memory. The model is based on the controversial assumption that context noise accounts for all forgetting in recognition. BCDMEM can be seen as an existence proof that a simple model taking only context noise into account, but not item noise, can explain a wide range of phenomena in recognition. More important for our purposes are the criticisms Dennis and Humphreys (2001) made with respect to the existence and boundary conditions of LLE and LSE in recognition. In particular, they claimed that when retention interval is long enough to allow proper context reinstatement, both LLE and LSE should disappear.

#### 1.6.2. SAC (Reder et al., 2000)

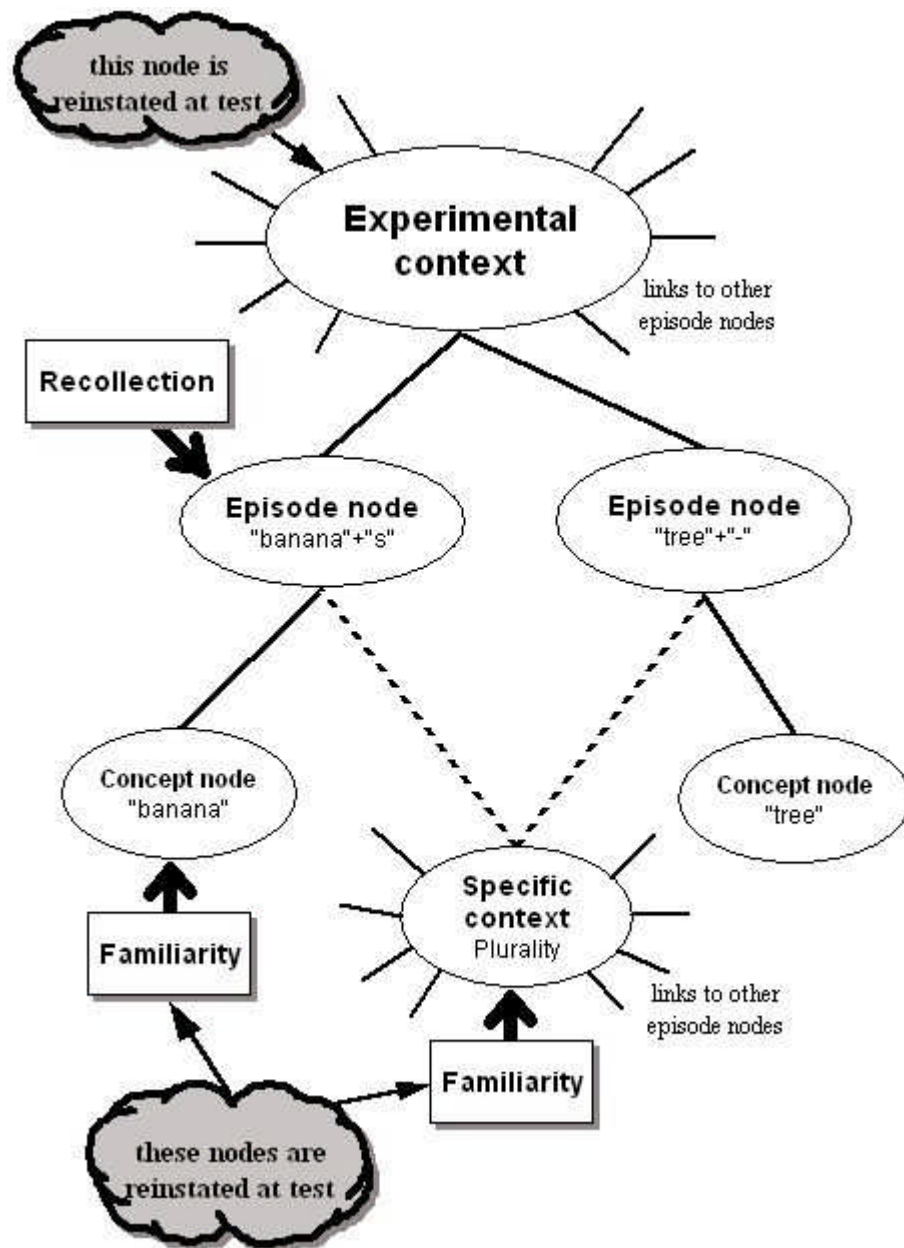
The growing body of evidence suggesting that recollection may contribute to item and associative recognition motivated the development of *mechanistic* dual-process models (as opposed to *measurement* models, which do not define how memories are represented). The Source of Activation Confusion model (SAC; Park, Reder, & Dickison, 2005; Reder et al., 2000) was one of the first mechanistic dual-process models developed to account for effects that were problematic to single-process models, such as the word-frequency effect, the registration-without-learning phenomenon and the dissociations obtained with the *Remember/Know* paradigm. Figure 1.3 illustrates the model in the context of

a switched-plurality experiment (for other instantiations of the SAC model, see Diana, Reder, Arndt, & Park, 2006).

In SAC, words are represented as nodes in an associative network. Each word is represented by a different node (*concept node*). All words are assumed to exist already in memory; they are not created during the experiment. At study, links are formed between a concept node and extra features associated with the word (e.g., the word's plurality status) during the study event. These extra features are also represented as separate nodes (*specific context nodes*). A word's plurality is not assumed to be represented together in the concept node because it represents a particular instance of the concept that may vary across events. The encoding of the word's plurality is assumed to be a probabilistic event, as plurality is a non-salient feature and thus may be missed during encoding. The association between a word (e.g., *banana*; concept node) and features present at study (e.g., *s*; specific context node) gives rise to an *episode node* (e.g., *banana + s*).

Whereas an episode node binds together one concept node to its specific context nodes, one specific context node may be bound to many episode nodes. This implements the common experimental design in which half of the words on a list are studied in their singular form and half are studied in their plural form. Episode nodes are bound together during the experimental session through their associations to the *experimental context node*, which represents the list-wide environment. Thus, the experimental context node is linked to all episode nodes created during a study session. Links emanating from the experimental context node can vary both in number and in strength. In sum, a network is created in SAC during study that simulates the storage of associations between words and environmental features and the relationships between those associations.





**Figure 1.3. Source of Activation Confusion (SAC) model.**

At study, links are formed between items (Concept Nodes; e.g., banana) and extra features associated with the item (Specific Context Node; e.g., the item's plurality). The association between an item and its specific context gives rise to an Episode Node. Solid lines represent deterministic links (i.e., a studied item is always connected to a node representing the current study episode). Dashed lines represent probabilistic links (i.e., sometimes the plurality feature is not encoded). Specific Context Node is shared with many episodes. Episode nodes are bound to the Experimental Context Node representing the list-wide environment. Thus the Experimental Context Node is linked to all Episode Nodes created during a study session. Links emanating from the Experimental Context Node can vary either in number (list-length manipulation) or in strength (list-strength manipulation). At test, an item is presented (e.g., *banana*) together with its plurality (e.g., *s*) and the context (e.g., current experiment). Activation spreads from those nodes to the rest of the network. If the level of activation in the corresponding Episode Node (e.g., *banana + s*) surpasses a threshold, a *Remember* response is produced. If the level of activation is below threshold in the Episode Node but above threshold in the corresponding Concept Node, a *Know* response is produced. Together, *Remember* and *Know* correspond to "old" responses. Adapted from Diana, Reder, Arndt and Park (2006). © American Psychological Society.

At test, a word is presented to the model together with its plurality and the experimental node, reinstating the original learning episode. For example, a cue representing the concept *banana*, its plurality *s* and the list in which the word was studied may be presented during a test trial. This cue reinstatement then triggers the spread of activation in the network from the reinstated nodes (concept + specific context + experimental context) to the rest of the network. For targets, activation spreads from their concept nodes and from the experimental context node to their corresponding episode nodes. The episode nodes may also receive activation from the specific context node if the words' pluralities were encoded at study. For switched-plurality lures, activation spreads from the concept node to their episode nodes (which were created at study because the concept was studied) and from the experimental context node, but no activation comes from the specific context node. As a result, activation in episode nodes for switched-plurality lures tends to be lower than activation in episode nodes for targets. For both targets and switched-plurality lures, if the summed activation in the episode node surpasses a response criterion, a *Remember* response is produced. If, on the other hand, activation is below criterion, decision is passed down to the concept node. If the activity in the concept node is higher than a (possibly different) response criterion, a *Know* response is produced. Note that both *Remember* and *Know* correspond to “old” responses. If neither episode node nor concept node surpass the response criteria, then a “new” response is produced.

Activation in SAC is modulated by the word's environmental frequency, the time since the word was last seen and the activation spread from other nodes in the network upon presentation of the word (or other linked words). The activation accrued to a given node, whether concept node, episode node, specific context node or experimental context node, is thus given by the sum:

$$A_{node} = B_{baseline} + \Delta A_{decay} + \Delta A_{spread} \quad (1.9)$$

where  $B_{baseline}$  is the node's baseline activation,  $\Delta A_{decay}$  is the activation lost due to decay and  $\Delta A_{spread}$  is the activation gained due to the spread of activation from other nodes. Baseline activation is given by  $B = B_w + \ln\left(c_N \sum_i t_i^{-d_N}\right)$ , where  $c_N$

and  $d_N$  are constants,  $t_i$  is the time since the words's last presentation and  $B_W$  is the base-level activation of the node ( $B_W$  is high for high-frequency words, low for low-frequency words and 0 for episode nodes). The change in activation due to decay is given by  $\Delta A = -\rho(A - B)$ , where  $\rho$  is a fixed parameter. This means that after each trial, the node's level of activation decreases by an amount that is proportional to the distance between baseline and current activation.

The activation received by a node  $r$  (receiving node) due to spread of activation from other nodes  $s$  (sending nodes) in the associative network is given by:

$$\Delta A_{spread} = \Delta A_r = \sum_s \left( A_s \times S_{s,r} / \sum_I S_{s,I} \right) \quad (1.10)$$

where  $A_s$  is the current activation of sending node  $s$ ,  $S_{s,r}$  is the strength of the link between nodes  $s$  and  $r$  and  $\sum_I S_{s,I}$  is the sum of the strengths of all links emanating from node  $s$ . Note that the amount of activation accrued to node  $r$  depends on how many nodes are connected to node  $r$  (i.e.,  $\sum_s$ ), on the current level of activation of each node  $s$  (i.e.,  $A_s$ ) and on the strength of the link between nodes  $s$  and  $r$  relative to the sum of the strengths of the links between node  $s$  and all the nodes  $I$  it is connected to (i.e.,  $S_{s,r} / \sum_I S_{s,I}$ ). Link strength is given by  $S_{s,r} = \ln(c_L \sum_i t_i^{-d_L})$ , where  $c_L$  and  $d_L$  are decay constants and  $t_i$  is the time since the  $i$ -th association between nodes  $s$  and  $r$  was formed.

The mapping between node activation and response probability is defined by assuming that node activation follows a normal distribution. For an *episode node*  $E$  with activation  $A_E$ , standard deviation  $\sigma_E$  and response criterion  $T_E$ , the probability of responding *Remember* is given by  $P(R) = \Phi[(A_E - T_E) / \sigma_E]$ , where  $\Phi(z)$  is the cumulative normal probability distribution (i.e., the area under the standard normal distribution to the left of  $z$ ). Likewise, for a *concept node*  $C$ , with activation  $A_C$ , standard deviation  $\sigma_C$  and response criterion  $T_C$ , the probability of responding *Know* (given that a word is not remembered) is given by  $P(K) = [1 - P(R)] \times \Phi[(A_C - T_C) / \sigma_C]$ . Note that  $P(\text{"old"}) = P(R) + P(K)$  and that responding *Know* to a lure is a false alarm. The probability of false-alarming

to a switched-plurality lure increases with the probability of recalling the item when its plurality has not been encoded and decreases with the probability of recalling the item when its plurality has been encoded (recall-to-reject). It is given by  $P(F_{SP\ lure}) = (1 - c)P(R) + [1 - P(R)]P(K)$ , where  $c$  is the probability that the plurality of a word is encoded at study (set to .5 in most SAC studies).

LLE and LSE are predicted in SAC as a consequence of the spread-of-activation process. The more nodes  $I$  a node  $s$  is connected to (i.e., the longer the study list) or the stronger the links between nodes  $I$  and node  $s$  relative to the strength between node  $r$  and node  $s$  (i.e., the stronger some items are in a *mixed* list;  $I \neq r$ ), the less activation is left to be spread from node  $s$  to node  $r$ . Less activation means lower probability of activation surpassing the response criterion and, therefore, lower probability of emitting a *Remember* response. Thus, according to SAC, list-length and list-strength manipulations should cause a reduction in the number of *Remember* responses. Consistently, Cary and Reder (2003, Exp. 3) found an LLE whereby *Remember* responses decreased and *Know* responses remained constant with increasing list length. Likewise, Diana and Reder (2005, Exp. 1) found a discrimination ( $d'$ ) LSE for *Remember* responses, whereby participants gave fewer *Remember* responses for weak items in the *mixed* list than in the *pure weak* list. Discrimination LSE for *Know* responses did not differ across lists. In both studies, results were fitted with the SAC model.

Interference in the SAC model, like in SAM and MINERVA2, occurs at retrieval. That is because each word is stored separately as a concept node. This is similar to the word representation adopted in BCDMEM, where words are represented as individual nodes in the input layer. Unlike BCDMEM, however, forgetting in SAC occurs due to interference within the same experimental context (i.e., competition for activation emanating from the experimental context node), not as a result of extra-experimental context interference.

SAC is not exactly a global matching model because the matching process does not take into account all items stored in memory. This is a consequence of SAC's architecture, which uses local nodes, and from the dynamics of its recognition

process, which uses a local match. At test, the concept node representing the test item is the only concept node directly activated; the concept nodes of other words may be activated only indirectly. Thus, although other stored traces may contribute to the recognition process, they do so indirectly through their connections via the experimental context node.

SAC can explain word-frequency effects (for normal and amnesic patients), LLE and LSE and time-course data (for a review, see Diana et al., 2006). In the case of the word-frequency effect, for example, SAC assumes that the hit-rate portion of the effect affects recollection and that the false-alarm portion affects familiarity. This assumption is consistent with behavioural and pharmacological studies (e.g., Hirshman et al., 2002; Joordens & Hockley, 2000; but see Malmberg, Zeelenberg et al., 2004). Baseline activation in concept nodes is higher for high-frequency than for low-frequency words. Thus, high-frequency lures are more likely to reach threshold in concept nodes than low-frequency lures. This accounts for the false-alarm portion of the word-frequency effect. The hit-rate portion of the effect is explained by the number of links emanating from the concept node. This number is larger for high-frequency than low-frequency words, implementing the idea that high-frequency words are linked to more episode nodes than low-frequency words (i.e., high-frequency words have been studied in more contexts). Targets presented at test are more likely to surpass the episode node threshold if they represent low-frequency than high-frequency words. This occurs because low-frequency words are associated with fewer episodes (i.e., fewer contexts) and, therefore, have more activation left to spread. This accounts for the hit-rate portion of the word-frequency effect.

In brief, SAC is a dual-process, abstract, network model that accounts for word-frequency effects, *Remember/Know* data and LLE and LSE in recognition. In particular, the model predicts that list-length and list-strength manipulations should harm recollection more than familiarity.

### 1.6.3. CLS (Norman & O'Reilly, 2003)

The Complementary Learning Systems model (CLS; Norman & O'Reilly, 2003) is a biologically plausible, dual-process, neural network whose architecture and

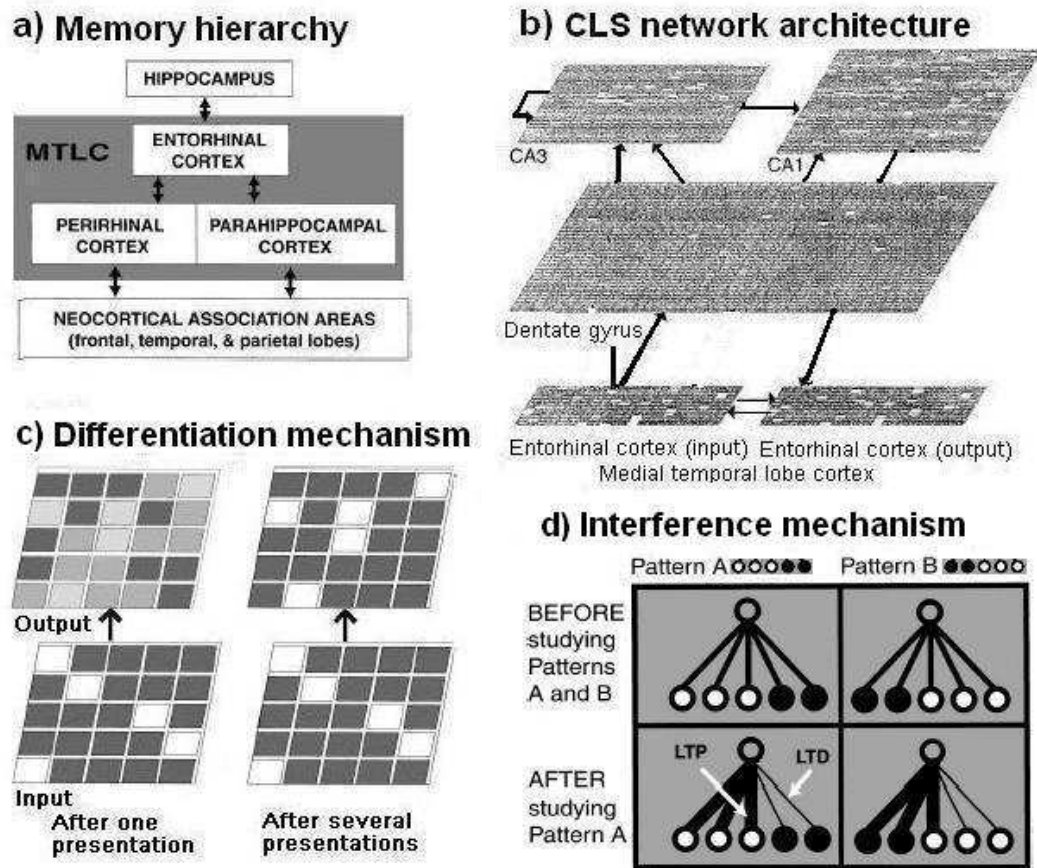
functionality map onto the psychological concepts of recollection and familiarity. Recollection is implemented in a module (the *hippocampal model*) that simulates the connectivity of the human hippocampus, an area essential for the learning and recall of declarative information (Aggleton & Brown, 1999; Eichenbaum, 2000). Familiarity is implemented in a module (the *cortical model*) that simulates cortical areas surrounding the hippocampus. These areas, jointly called the Medial Temporal Lobe Cortex (MTLC), have been implicated in recognition without the recall of details (Aggleton & Brown, 1999; Bowles et al., 2007).<sup>13</sup>

Together, the hippocampal and cortical models instantiate the tenets of the Complementary Learning Systems framework in recognition memory (McClelland, McNaughton, & O'Reilly, 1995). According to this framework, the hippocampus has evolved a specialised network capable of rapidly memorising specific events, whereas the MTLC evolved a specialised network for slowly learning the statistical regularities of the environment. Learning occurs through Hebbian Long-Term Potentiation (LTP: the connection strength between two neural units is increased if they are both active at the same time) and Long-Term Depression (LTD: the connection strength between two neural units is decreased if the receiving unit is active but the sending unit is not).<sup>14</sup> The hippocampal model is able to learn quickly without suffering catastrophic interference because it assigns relatively distinct representations to input stimuli. By contrast, the cortical model is able to learn regularities (e.g., prototypes) by assigning relatively similar representations to similar stimuli; it can thus generalise its activation to novel stimuli based on previous experience with similar stimuli. Figure 1.4 (a,b) depicts the overall network architecture of the CLS model.

---

<sup>13</sup> The claim that the hippocampus is implicated only in recollection is quite controversial. Studies have found spared recall after hippocampal lesions (Mayes et al., 2001), suggesting that an intact hippocampus is not necessary to elicit recollection, and equal impairment of recall and recognition after hippocampal lesions (Manns et al., 2003), suggesting that the hippocampus is important for both recollection and familiarity (see Squire, Wixted, & Clark, 2007, for a review).

<sup>14</sup> For a review on neural networks and Hebbian learning, see O'Reilly and Munakata (2000).



**Figure 1.4. Complementary Learning Systems (CLS) model.**

(a) Macro-architecture. Highly processed input coming from specialised brain areas converge into the medial temporal lobe cortex (MTLC). MTLC processes the signal and passes it on to the hippocampus, which processes the signal and sends it back to MTLC. MTLC then sends the signal back to the specialised brain areas. (b) Micro-architecture. Input activation is passed on to the entorhinal cortex (the cortical model) which assigns similar representations to similar items. The average activation of the top 10% units is the measure of familiarity. Information is then passed on to the hippocampal model (dentate gyrus, areas CA3 and CA1). The dentate gyrus takes the overlapping input pattern and turns it into a non-overlapping, sparse pattern. Area CA3 takes that pattern and re-represents it in a way that allows for the later reconstruction of the entire pattern from portions of it. This process, called pattern completion, is akin to cued recall and it is possible due to CA3's recurrent connections. Recollected patterns in CA3 are then translated back into overlapping representations in CA1, which then sends the pattern to the entorhinal cortex. The pattern presented in the input layer of entorhinal cortex is then compared to the pattern recollected on the output layer of the entorhinal cortex. The model adopts a recall-to-reject rule, whereby it responds “new” if there is any mismatch between input features and recollected features. Positive LLE and LSE are predicted in the hippocampal model because its activation distribution has room to decrease and the lure activation distribution is already at floor. (c) Differentiation mechanism in MTLC. New test items weakly activate many units in MTLC, whereas old test items strongly activate few units. This differentiation process, implemented through Hebbian learning and lateral inhibition, underlies the prediction of a null LLE and null LSE in the MTLC model, as both target and lure activation distributions decrease as a result of interference from other items. (d) Interference mechanism. Interference increases the weights to features shared by many items on the study list (LTP) and decreases weights to discriminative features of studied items and lures (LTD). In CLS, length and strength manipulations produce similar levels of interference. In all figures, high neural activity is represented in white colour. Adapted from Norman and O'Reilly (2003). © American Psychological Association.

The signals produced by the hippocampal and cortical models have different operating characteristics. In the hippocampal model, old test items may trigger activation (recollection) of stored traces but lure items rarely do so. This is a consequence of *pattern separation* (i.e., assignment of distinct representations to different items, regardless of similarity). In the cortical model, however, both old and lure items may trigger activation (familiarity). This follows from the MTLC's assignment of *overlapping* representations to different items. The architectural differences between the models yield signals with different properties: the hippocampal model behaves much like a threshold process, whereas the cortical model behaves like a standard signal detection process.

In the cortical model, new items weakly activate a large number of units, whereas old items strongly activate a small number of units (Figure 1.4, c). This *differentiation* mechanism follows from the joint operation of Hebbian learning and lateral inhibition: study presentation strengthens the connections between input units and the hidden units in MTLC through Hebbian learning; the winning units in MTLC then enhance signal contrast by blocking the weaker units within the same layer through a process lateral inhibition. Thus, test lures are less likely to accidentally activate a strong trace than a weak trace. This differentiation mechanism contrasts with REM's in that strong items activate more, not fewer, features in REM; that is because in REM encoding more features increases the probability of a mismatch by a lure, and mismatches decrease the odds signal.

*Familiarity* in the cortical model is given by the average activation of the top 10% units in MTLC produced as a result of the presentation of a test item. The *response criterion* is set by averaging the familiarity signals from studied and lure items and then placing the criterion between the corresponding means. An item is called "old" if its activation is above criterion.

In the hippocampal model (dentate gyrus, area CA3 and area CA1), the inputs from MTLC (entorhinal cortex in Figure 1.4, b) are transformed and stored in a way that allows their subsequent recollection upon presentation of the test item, even if the test item represents a degraded version of the original study item. The dentate gyrus takes as input the MTLC patterns and assigns them a sparse and



non-overlapping representation (*pattern-separation*).<sup>15</sup> Area CA3 then takes those sparse patterns and re-represents them in an auto-associative network capable of *pattern-completion* (i.e., CA3 can reconstruct a complete version of a previously stored pattern from a partial version of it; this is possible due to its recurrent connectivity). This process, akin to cued recall, implements recollection in the hippocampal model. Area CA1 then translates the sparse, non-overlapping representation reconstructed in CA3 back to its original overlapping form and sends it to the entorhinal cortex. The entorhinal cortex thus serves both as an input layer, presenting overlapping patterns to the hippocampus, and as an output layer, receiving the recollected patterns from the hippocampus.

*Recollection* in the hippocampal model is given by the difference in the number of matching and mismatching features between test pattern and recollected pattern. *Recall-to-reject* is defined as a rule: respond “new” if the test item yields any recollected mismatch; respond “old” otherwise. The rule is plausible because any mismatch between test and recollected patterns is strong indication that the item was not studied. Recent neuroimaging studies provided some support for the properties implemented in the hippocampal model, in particular, its ability to carry out pattern separation (Kirwan & Stark, 2007) and its role as a match-mismatch detector (Kumaran & Maguire, 2007a, 2007b).

The CLS model is a global matching model. Recognition of a test item is based on the joint contribution of all stored items. Like TODAM, but unlike SAM, MINERVA2 and SAC, traces in CLS are stored in a composite vector, comprising the hidden layers of the hippocampal and cortical models. Consequently, interference effects in CLS occur at study, not at retrieval. The overall effect of interference on a given trace due to additional presentations of an item or to additional items is to decrease weights to discriminative features of studied items and lures (i.e., features that distinguish between similar items are decremented) and to increase weights to prototypical features (i.e., features that are shared by many items on the study list are incremented). This interference mechanism, implemented through the biologically plausible processes of LTP

---

<sup>15</sup> A representation is *sparse* if only a few of its features are active. A representation is *non-overlapping* if the probability that it shares a feature with another representation is small.

and LTD, is the same in both the cortical and the hippocampal modules (see Figure 1.4, d). The impact of interference, however, is different across modules.

The differential impact of interference across modules occurs mostly as a result of their architectural differences. The main prediction of the CLS model with respect to LLE and LSE is that, for some parameter values<sup>16</sup>, the hippocampal module is more strongly affected by length and strength manipulations than the cortical module. This is consistent with evidence showing that an LSE emerges when recollection is likely to operate but it does not emerge when familiarity is likely to operate (Norman, 2002; Norman et al., 2008; Verde & Rotello, 2004).

At first, this seems counterintuitive given the ability of the hippocampal model to carry out pattern separation. Yet some representations do overlap in area CA3 (Norman & O'Reilly, 2003, p. 629). Consequently, the strengthening of some stored items may result in the shrinkage of weights to discriminative features of other traces stored in CA3. The end result is that some weak items may not be able to trigger recollection at test due to the smaller activation produced by their smaller connection weights. This accounts for the decrease in hits in the hippocampal model in the weak condition. The false alarms, on the other hand, have no room to decrease further because they are already at floor (lure items rarely trigger recollection). Thus, the overall discrimination for weak items in *mixed* lists decreases in the hippocampal model.

In the cortical module, on the other hand, the strengthening of some stored traces results in the decrease in weights to discriminative features of both stored targets and lures (although lures are not present in the studied list, they are nonetheless represented in the memory system and their representation can become stronger or weaker). The simultaneous decrease in the weights of targets and lures occurs because the representations in the cortical model are overlapping. Thus, if the activation of a test item is reduced, the activation of a similar item (which activates shared features through the same weights) is also reduced. Crucially,

---

<sup>16</sup> The most important parameter in the derivation of these predictions is the *average input overlap* (i.e., similarity between study items). The prediction of higher interference in the hippocampal model depends on input overlap being low. This can be experimentally achieved by using lists of unrelated items. All the experiments reported in this thesis abide by this restriction.

the activation of lures is not at floor in this model, thereby allowing false alarms to decrease. As a result, both target (hits) and lure (false alarm) distributions decrease with interference, resulting in no net difference in discriminability.

The CLS model predicts an LLE for the same reason it predicts an LSE: adding more items causes similar weight changes as strengthening some items (Norman & O'Reilly, 2003, p. 632). Although the CLS model predicts both LLE and LSE in recognition, there are boundary conditions on the predicted effects. One variable modulating the size of the effects is study-test lag. Longer lags, which are equivalent to longer retention intervals in a retroactive design, should produce less pronounced or even null LLEs and LSEs compared to shorter lags. Intuitively, there should be less interference of strong items over weak items if the strong items become weaker, and this can happen when retention interval is long (Shepard, 1967; Strong, 1913). Norman and O'Reilly (2003, p. 632) simulated the effects of retention interval by using dynamic weights. Unlike the static weights used in their previous simulations, dynamic weights change over time. They reach peak value at each presentation of an item and then decay exponentially with time. As a consequence, weights are larger at shorter than at longer retention intervals because they do not have time to decay. Larger weights produce stronger interference effects due to their greater disruptive influence on activation patterns. Therefore, according to the CLS model, larger LLE and LSE are predicted at short than at long retention intervals.

CLS also accounts for previous results that were problematic for single-process models, such as the non-monotonic false-alarm curves in item and associative recognition (see 1.5.3) and the forced-choice advantage for non-overlapping pairs in associative recognition. Non-monotonic false alarms (Gronlund & Ratcliff, 1989; Hintzman & Curran, 1994) can be explained by the model's architecture. False alarms initially rise because activation spreads first to the cortical model, which is sensitive to item similarity, thereby producing high levels of familiarity to both targets and similar lures. False alarms eventually fall because activation then spreads to the hippocampal model, which is less sensitive to item similarity, thereby producing reliable mismatches that can be used to reject similar lures.

In associative recognition, the advantage for non-overlapping pairs refers to the fact that, in a forced-choice task, participants are better at choosing targets (AB) among distractors that do not share any of the target's items (CF, GJ; non-overlapping pairs) than at choosing targets among distractors that do share one of the target's items (AD, AF; overlapping pairs) (Clark, Hori, & Callan, 1993). The result contradicted several global matching models of the time (e.g., MINERVA2, TODAM). The models predicted an advantage for overlapping pairs because the familiarities of similar test items are correlated and thus the variance of the difference between targets and distractors should be smaller.<sup>17</sup> CLS accounts for the non-overlapping advantage because the highly diagnostic recall-to-reject process has more opportunities to operate with non-overlapping than with overlapping pairs. The former provides 6 opportunities for the model to try and recall the pair (A,B,C,F,G,J); the former provides 4 opportunities only (A,B,D,F).

In sum, the CLS model is a dual-process, biologically plausible, connectionist model that accounts for LLE and LSE in recognition. In particular, the model predicts that list-length and list-strength manipulations should harm recollection more than familiarity and that retention interval should modulate those effects.

## 1.7. Aims of the thesis

The findings of a null LLE (Dennis & Humphreys, 2001) and a positive LSE (Norman, 2002) highlight the uncertainty over the empirical status of list-length and list-strength effects in recognition and call for further investigation of the variables that determine the occurrence of these effects. The study of LLE and LSE is particularly important because their prediction follows from core assumptions of early global-matching models (SAM, MINERVA2, TODAM) and from certain parameter settings of more recent single-process (REM, BCDMEM) and dual-process models (SAC, CLS). In the following, we present the main empirical and theoretical objectives of the work reported in this thesis.

---

<sup>17</sup> Formally, the variance of the difference between targets ( $T$ ) and distractors ( $D$ ) is given by  $\sigma_{T-D}^2 = \sigma_D^2 + \sigma_T^2 - 2\rho\sigma_D\sigma_T$ , where  $\sigma^2$  is variance and  $\rho$  is the correlation of  $T$  and  $D$ . If the features of the test pairs overlap, then  $\rho$  increases and  $\sigma_{T-D}^2$  decreases, thereby boosting  $d'$ .

### 1.7.1. Empirical objectives

In this thesis, we report several item recognition experiments in which list length, list strength, encoding task, lure relatedness, and retention interval were manipulated within and between participants. We report two series of experiments: the first series (Experiments 1, 2, 3 and 4) followed the design used by Dennis and Humphreys (2001) and is described in Chapter 3; the second series (Experiments 5a, 5b, 6 and 7) followed the design of Norman (2002) and is described in Chapter 4. The eight experiments reported in this thesis were carried out aiming to achieve five empirical aims.

The first empirical aim of this thesis is to test the hypothesis that Dennis and Humphreys' (2001) null LLE and null LSE were obtained as a result of low recollection rates at test. List-length and list-strength manipulations may impair recollection more than familiarity (Diana & Reder, 2005; Norman & O'Reilly, 2003). Accordingly, previous null results could be explained by a relatively small contribution of recollection processes at test. If targets and distractors are highly dissimilar, familiarity alone may be a reliable basis for recognition judgments. However, if targets and lures are similar, familiarity alone may not be diagnostic (e.g., one needs to recollect seeing *banana* to reject lure *bananas*). In Dennis and Humphreys' (2001) experiments, targets and distractors were dissimilar, which raises the possibility that responses in those experiment were mostly based on familiarity. To investigate this possibility, we varied target-lure similarity in Experiments 1 to 4 (to manipulate the likelihood of recall-to-reject) and assessed its impact on length and strength effects, while keeping the design as similar as possible to Dennis and Humphreys' (2001) design. We also manipulated target-lure similarity in Experiment 7, while keeping the design as similar as possible to Norman's (2002) design.

The second empirical aim of this thesis is to evaluate the role played by possible confounds present in previous studies. In Experiments 1 to 4, we tested whether Dennis and Humphreys' (2001) null LLE and null LSE were a consequence of participants' use of covert rehearsal strategies at study. Because encoding time in most studies has been fixed (and long), rather than self-paced (and short),

participants could have used some of the encoding time to rehearse previously presented items. This is particularly true for strength manipulations, where study items are repeated. After one repetition, participants may decide that they have already learned the item and use the remaining study time to practice weaker items. To the extent that rehearsal occurs, it would work towards reducing the possibility of finding an effect. To address this issue, we used a self-paced encoding task in Experiments 1 to 4, while keeping other design features as close as possible to Dennis and Humphreys' (2001). Moreover, in Experiments 5 to 7, we used short encoding times, following Norman's (2002) design, to reduce the use of covert rehearsal. Finally, the inclusion of a list-length manipulation in our studies also addresses a potential confound present in Norman's (2002) study. Norman (2002) compared weak short lists with long strong lists. Because repeating study items entails a longer list, length of list presentation and list strength were confounded. We control for that by having both a list-length and a matched list-strength manipulation in all studies reported here.

The third empirical aim of this thesis is to assess the importance of the encoding task in the production of LLE and LSE. Dennis and Humphreys (2001) used a pleasantness rating task at study, whereas Norman used a size judgment task, in which participants had to decide whether a typical exemplar of a study word (e.g. *banana*) would fit into a banker's box present in the experimental room. As argued by Norman (2002), the purpose of the size judgment task was to increase the chances of trace overlap, as all words would presumably be encoded with a common referent (the banker's box), thus increasing memory interference. In the pleasantness rating task, by contrast, each word may have been encoded with a different referent, reducing the chances of interference. To test the possibility that the encoding task may have been responsible for the LSE observed by Norman (2002) and the null LSE reported by Dennis and Humphreys (2001), encoding task (size vs. pleasantness judgment) was manipulated in Experiments 1 and 3.

The fourth empirical aim of this thesis is to assess the role of retention interval on the size of LLE and LSE. *Retention interval* is defined here as the amount of time elapsed between presentation of the last study word and presentation of the first test word. It should not be confused with study-test lag, which is the average

amount of time between the presentation of an item and its subsequent test. Retention interval has been identified as a potentially relevant factor in LLE and LSE. Dennis and Humphreys (2001) argued that recognition of items from the beginning of the list should be impaired if the test is taken immediately after study. The main hypothesis is that LLE and LSE should increase when retention interval decreases. This hypothesis follows from several memory models. Nevertheless, the role of retention interval in LLE and LSE has not been tested yet. Cary and Reder (2003, Experiment 3) found an LLE after controlling for the confounds identified by Dennis and Humphreys (2001), including retention interval, but they did not manipulate retention interval itself, keeping it constant at 120 s. Similarly, Norman (2002) obtained an LSE using a 120-s interval in his strength condition, without manipulating the retention interval. We addressed this issue by varying the retention interval in the long and strong lists (0 s vs. 180 s in Experiments 1 to 4; 10 s vs. 120 s in Experiments 5 to 7). The manipulation was carried out between participants (Experiments 2 to 4) and within participants (Experiments 5 to 7). In Experiments 5 to 7, we also manipulated whether unrelated lures would be present at test. The aim was to increase recollection-based discriminability in order to facilitate the detection of differences among list types. Discriminability was reported to be better when only related lures are used at test ('without new' condition) compared to when both related and unrelated lures are used ('with new' condition) (Heathcote, Raymond, & Dunn, 2006).

The final empirical aim of this thesis is to assess the impact of the number of repetitions on LSE and the impact of long-to-short list-length ratios on LLE. Although previous studies (LLE: Cary & Reder, 2003; LSE: Norman, 1999) have found suggestive evidence that more repetitions yield larger LSEs and longer lists yield larger LLEs, the results were not conclusive. Norman (1999) found that increasing the number of presentations of strong items from three to six in *mixed* lists resulted in an increase in the size of the list-strength effect. However, Norman (1999) based his conclusions on SDT measures derived from single-point data (i.e., old–new recognition task). Single-point discriminability is subject to distortions when the response criterion varies across conditions (see 2.2.3, for a discussion of discriminability measures), and strength manipulations are usually accompanied by criterion shifts (e.g., Hirshman, 1995). Results based

on single-point data may thus confound discriminability differences with criterion differences. We addressed this issue by collecting confidence ratings at study, which allows the analyses to take criterion shifts into account. Cary and Reder (2003, Exp. 3) found a list-length effect with the same controls used by Dennis and Humphreys (2001, Exp. 1 and 2). One possible difference is that the former used a long-to-short ratio of 4:1, whereas the latter used 3:1 and 2:1 ratios. In order to test whether the difference in the size of the manipulation can explain the discrepant results, we manipulated the length ratios from small (2:1) to large (3.5:1). We assessed the impact of both the number of repetitions on LSE and the long-to-short list-length ratios on LLE by comparing the results of Experiments 2, 3 and 4 and the results of Experiments 5*b* and 6.

To summarise, our main goal in this thesis is to investigate the boundary conditions behind the list-length and list-strength effects. We pursued this goal by closely comparing the experiments carried out by Dennis and Humphreys (2001), where null LLE and null LSE were found, and Norman (2002), where a positive LSE was found. Several variables were manipulated with the objective of isolating the factors that seem essential in the production and modulation of LLE and LSE. The experiments reported here differed from Norman's (2002) in that we varied list length. The experiments here also differed from Dennis and Humphreys' (2001) in that we varied target-lure similarity. Finally, the experiments reported here differed from both Dennis and Humphreys' (2001) and Norman's (2002) in that we varied encoding time (self-paced vs. 1.15 s), encoding task (size judgment vs. pleasantness judgment), retention interval (short and long) and manipulation strength [3 presentations (3x) vs. 6 (6x) in list-strength manipulations; 2:1 length ratio vs. 3.5:1 in list-length manipulations].

### 1.7.2. Theoretical objectives

The experiments reported in this thesis address questions of theoretical interest. In particular, the experiments test convergent and divergent predictions from state-of-the-art recognition models, such as BCDMEM, REM, SAC and CLS.

The first theoretical objective is to test the role of recollection on LLE and LSE. Both SAC and CLS predict, for some parameter values, that recollection should



be more impaired than familiarity when list-length and list-strength are manipulated. We test this prediction by manipulating target-lure similarity. Any results showing a dissociation, whereby discrimination between targets and switched-plurality lures is more impaired than discrimination between targets and unrelated lures, would constitute evidence in favour of SAC and CLS. On the other hand, such a result would provide evidence against BCDMEM. BCDMEM can predict LLE and LSE through its context reinstatement parameter but it cannot predict differential effects on lure types.<sup>18</sup>

The second theoretical objective is to test the modulatory role of retention interval on LLE and LSE. Both BCDMEM and CLS (and possibly SAC<sup>19</sup>) predict larger effects when retention interval is short. BCDMEM predicts the effects as a consequence of poor reinstatement of the study context at test; CLS predicts the effects as a result of the higher values of time-dependent weights in its network. BCDMEM assigns no causal role to interference from other items on the list, whereas CLS does. To test this prediction, which is shared by both models, we varied retention interval. To differentiate between the models, we also varied lure type. BCDMEM predicts equal interference for both unrelated and switched-plurality lures, whereas CLS predicts a stronger effect of retention interval for switched-plurality lures. Any result showing a differential LLE and LSE between unrelated and switched-plurality lures would thus be evidence for CLS and against BCDMEM. By contrast, a result showing similar changes in LLE and LSE across lure types and retention intervals would constitute evidence for BCDMEM and against CLS.

The third theoretical objective is to assess the impact of stronger list-strength manipulations on the magnitudes of the LSE. REM predicts null LSEs regardless

---

<sup>18</sup> Dissociations across lure types could also be interpreted as evidence against REM. In associative recognition, REM correctly predicts an LLE for new pairs but incorrectly predicts no LLE for rearranged pairs due to the larger weight given to matches compared to mismatches in REM (Criss & McClelland, 2006, Simulation 2). Yet we cannot make confident predictions about item recognition because the relevant derivations concerning LLE and LSE across lures have not been reported in REM's publications (Malmberg, Holden et al., 2004; Shiffrin & Steyvers, 1997).

<sup>19</sup> SAC could also predict larger LLEs and LSEs for short retention intervals, as the network's link strengths decay with time. This means that, at shorter intervals, link strengths would be larger and more disruptive. However, we refrain from making strong claims about the effects of retention interval and number of repetitions (see next paragraph) on LSE, since these predictions were not explicitly derived in SAC's publications (e.g., Diana & Reder, 2005; Reder et al., 2000).

of the size of the strength manipulation, whereas CLS predicts ever increasing LSEs. REM predicts null LSEs because additional strengthening will make strong items even more differentiated compared to weak items; the odds of an accidental match would, therefore, approach zero. In CLS, by contrast, additional strengthening would cause additional disruption of stored traces; as a result, discriminability should approach zero with increasing strength interference.

The final theoretical objective is to compare the relative sizes of LLE and LSE. Both SAC and CLS predict that LLE should be either equal or larger than LSE. In SAC, this follows from the way activation spreads in the model. Adding more words causes a greater decrease in activation than strengthening some words. In the former, the decrease in activation is linear (i.e.,  $\sum_I S_{s,I}$ ); in the latter, it is logarithmic [i.e.,  $\ln(c_L \sum_i t_i^{-d_L})$ ; see Equations 1.9 and 1.10]. In CLS, the length and strength manipulations induce weight changes of similar magnitude. In another version of CLS, however, where weight values change over time, the model predicts a length-strength dissociation, whereby an LLE is produced but not an LSE (Norman & O'Reilly, 2003, p. 632). Any result showing the opposite pattern (i.e., LSE larger than LLE) would provide evidence against both models.

In sum, the experiments reported in this thesis address questions of theoretical relevance with the potential of differentiating between alternative models. In the next chapter we describe the analytical tools used to interpret the results reported in this thesis. We then describe our first four experiments in Chapter 3 and the remaining four experiments in Chapter 4. In Chapter 5, we conclude by discussing both the empirical and theoretical implications of these results.

## Chapter 2. General Methodology

### 2.1. Introduction

In this chapter, we describe the methodology used to analyse the results presented in Chapters 3 and 4. Data was analysed with relatively assumption-free measures of performance (hits, false alarms and response times) and assumption-dependent measures of performance (sensitivity  $A_z$  and bias  $c_a$ ).

The *hit rate* is the proportion of *old* trials in which the participant correctly responded “old”; each such event is called a *hit*. The *false-alarm rate* is the proportion of *new* trials in which the participant incorrectly responded “old”; each such error is a *false alarm*. The *response time* is the period of time taken between the onset of a study or test item and the entry of a response. Hit and false-alarm rates for each experiment and condition are provided in Appendix 1; response times are described and analysed in Appendix 2.

Although hits and false alarms provide the most direct measures of performance, they have to be interpreted in the context of a decision mechanism. That is because people’s responses to a given memory cue may depend on factors other than the simple feeling of familiarity elicited by the cue. For example, confirming that a particular person present in a room was previously seen in a particular place may require more subjective evidence when this judgement is made among strangers in the dock than among friends in a pub. Although the feeling elicited by seeing the person may be the same in both cases, the amount of subjective evidence required to say “yes, I have seen this person in that place” will probably be higher in the first case. Thus, the external response given by an individual when faced with a memory cue may vary from situation to situation even though the internal memory feeling is the same. What is needed, therefore, is a way of disentangling the strength of the internal feeling of memory from the external factors influencing the way the memory is reported. Signal Detection Theory (SDT) provides us with one way of doing just that.

In the following, we give a brief overview of the assumptions underlying SDT. Next, we describe how it is possible to obtain separate measures of *sensitivity* (which relates to the strength of memory) and *criterion placement* (which relates to the amount of evidence deemed sufficient to output a positive response). Finally, we describe in some detail how those measures were estimated from the hit and false-alarm data produced in our experiments.

## 2.2. Signal Detection Theory

Signal Detection Theory (SDT) was used in this thesis to derive measures of sensitivity ( $A_z$ ) and response bias ( $c_d$ ). It is important to emphasise at the outset that our use of SDT is motivated mainly by pragmatic reasons. SDT models are used here because they tend to provide good fits to Receiver Operating Characteristic (ROC) curves obtained from hits and false alarms. Sensitivity  $A_z$ , for example, is an estimate of the area under the ROC curve obtained from the parameters of an (un)equal-variance SDT model fit to recognition data. If a model is able to fit the data well, then estimates derived from this model provide valid descriptions of the data. Note that one does not have to accept SDT as a *psychological theory* of human memory in order to use it as a *descriptive theory*.

SDT, as it is applied to recognition memory, is based on three assumptions: *i*) the evidence about whether a test item represents a *target* or a *distractor* can be summarised by a single number; *ii*) the evidence signal elicited by the test item is subject to random variation; *iii*) the choice of response (“old”, “new”) is made by applying a decision criterion to the magnitude of the evidence (Wickens, 2002).

The first assumption is supported by behavioural, neurophysiological and neuroimaging findings suggesting that familiarity decisions are based on a continuous internal signal (*familiarity*; see 2.2.1). The second assumption is supported by the fact that encoding of study items is subject to variations in participant’s attention or item distinctiveness; responses to the same items in the same conditions may vary from situation to situation. The third assumption is supported by the observation that factors other than the memorability of the item

itself, such as pay-offs (i.e., high vs. low penalties for making a false alarm), may influence the recognition decision.

### 2.2.1. Familiarity distribution

#### Familiarity signal

The idea that recognition decisions are based on a continuous variable is supported by early behavioural studies which found a smooth and monotonic relationship between the probability that an item was studied and participant's memory judgements along a confidence scale (e.g., Lockhart & Murdock, 1970; Mickes, Wixted, & Wais, 2007). Neuroimaging studies also provided support for the continuity hypothesis by showing that activity in the perirhinal cortex, an area adjacent to the hippocampus in the medial temporal lobe, is modulated at test by the item's status (*target* vs. *lure*) and that this activity tracks the perceived levels of confidence reported by the participants (Gonsalves et al., 2005; Henson et al., 2003). The idea that a signal in the perirhinal cortex could work as an index of memory strength is further supported by single-cell recordings in animal studies showing that changes in perirhinal neuronal firing, which are dependent on the relative novelty of the test item, occur after a single encounter with an item, can emerge as early as 75 ms after stimulus onset and can last for over 24 h (M. W. Brown & Aggleton, 2001; M. W. Brown & Xiang, 1998).

Signal detection theory is agnostic with respect to trace representation, memory encoding and retrieval processes. All SDT assumes is that a continuous and variable signal is elicited at test. In order to flesh out the processes producing the familiarity signal, process models such as the ones described in Chapter 1 are necessary. Note that, because SDT does not commit to a particular process model (all it cares is that there is a continuous evidence signal), the theory does not entail a *single process* in the sense used by process models, such as SAM or REM. The use of the term *familiarity* in the context of SDT is not to be confused with the term familiarity as it is used in single- versus dual-process controversies. In fact, it is possible to construct a dual-process SDT model simply by combining two sources of evidence, one coming from a recollection process and one coming from a familiarity process (Wixted, 2007; Wixted & Stretch, 2004). In sum,

extant evidence suggests that familiarity can be represented as a continuous random variable. SDT use this variable to model recognition data, regardless of the specifics of the underlying processes generating the signal.

### Equal-variance distributions

The simplest version of an SDT model applied to recognition memory involves two equal-variance distributions, one representing the *target* items and the other representing the *lures* or *distractors*, and a decision criterion placed somewhere along the familiarity continuum. A familiarity signal is assumed to be elicited in response to a test cue. If the signal is large enough to exceed the criterion, an “old” response is produced; otherwise, a “new” response is produced.

Separate *target* and *lure* distributions originate from recent exposure to a set of study items. Initially, it is assumed that a pool of items (lexical/semantic units) is available in memory. These items, which can be thought to represent a person’s word knowledge, produce familiarity signals whose values revolve around a common mean. Some of these items have higher levels of familiarity because they were more recently seen outside the laboratory; other items are less familiar. During the study phase of a memory experiment, some items in the initial pool are strengthened. Some studied items have their initial strengths incremented by a large amount, whereas others items have their strengths incremented by a small amount. This occurs due to variability in the encoding process.

The end result of studying some items is the separation of the original familiarity distributions into two distributions, with the mean familiarity of the recently studied items shifted to a higher level. However, because some of the unstudied items were originally highly active and because some of the studied items may have not been encoded properly, the distribution of familiarities for studied and unstudied items tends to overlap: some studied items are not later recognised during the test (*misses*) and some unstudied items are falsely recognised (*false alarms*). Thus, errors tend to occur during recognition, and accuracy depends on both the difference between the distribution means and on their degree of overlap. Figure 1.1 illustrates this process. The horizontal axis represents the

familiarity dimension along which the response to a test items is measured; the height of the distributions indicates how likely that familiarity value is to occur.

An equal-variance SDT model is defined by three parameters (assuming that the distractor distribution is centred around zero;  $\mu_D = 0$ ): (i)  $\mu_T$ , the mean of the *target* distribution; (ii)  $\sigma_D = \sigma_T = \sigma$ , the common standard deviation and (iii)  $X$ , the value on the familiarity continuum that separates “old” from “new” responses (criterion). If it is assumed that  $\sigma_D = 1$ , then an equal-variance SDT model can be characterised by two parameters ( $\mu_T, X$ ), estimated from hits and false alarms.

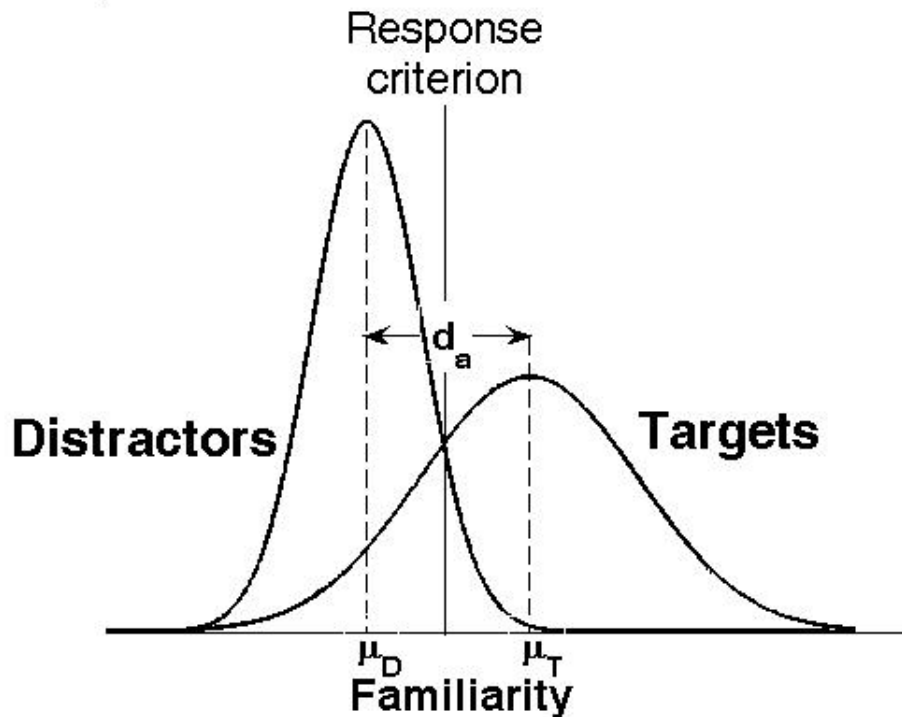
Although simple and useful for illustrative purposes, equal-variance models are rarely used in practice. There are at least two reasons for that. First, the model assumes that each study item shifts its original familiarity level from the initial pool by a fixed amount, which is an implausible assumption given the variability inherently present at encoding (Nelson, 2003). Second, equal-variance models do not provide good fits to experimental data; the variance of *target* distributions, estimated from empirical ROC curves, is generally 1.25 times larger than the variance of the *distractor* distributions (see Wixted, 2007, for a review). Thus, equal-variance SDT models are unlikely to provide reasonable fits to our data.

### Unequal-variance distributions

An unequal-variance SDT model is defined by four parameters (assuming  $\mu_D = 0$ ): (i)  $\mu_T$ , the mean of the *target* distribution; (ii)  $\sigma_D$ , the standard deviation of the *distractor* distribution; (iii)  $\sigma_T$ , the standard deviation of the *target* distribution and (iv)  $X$ , the familiarity value separating “old” from “new” responses (criterion). It is usually assumed that  $\sigma_D = 1$ , since only the ratio of standard deviations is of interest. Thus, in most situations, unequal-variance SDT models are characterised by three parameters ( $\mu_T, \sigma_T, X$ ). Figure 2.1 illustrates the model.

As mentioned in the previous session, most evidence points to unequal-variance SDT models as better descriptors of recognition data (e.g., Ratcliff et al., 1992). Moreover, the assumption of identical familiarity increments for each study item, necessary to keep the variances of *targets* and *lures* the same, is implausible. A more direct test of the unequal-variance assumption was carried out recently by

Mickes et al. (2007) who simply asked participants to rate on a wide scale how strong they felt their memory was with respect to each test item (the scales in their study ranged from 1 to 20 in Exp. 1 and from 1 to 99 in Exp. 2).



**Figure 2.1. Unequal variance SDT model and sensitivity measure.**

The familiarity signals elicited by the presentation of targets and lures at test follow distributions that can vary in their means and variances. When the variances are equal, discriminability ( $d'$ ) can be measured as the difference between the means in the common standard deviation units. When the variances are different, discrimination ( $d_a$ ) can be measured as the difference between the means in standard deviation units of the average (root mean square) of the two variances.

Mickes et al.'s (2007) results showed that the distribution of responses along the rating scales agreed with the assumptions of an unequal-variance SDT. First, the ratings' distributions for *targets* and *lures* overlapped and the mean of the *target* distribution was higher than the mean of the *lure* distribution. Second, accuracy rose continuously as the distance from the indifference point increased. Third, the variance of the ratings data, estimated for *targets* and *lures*, showed that the *target* ratings' distribution was about 1.25 larger than the *lure* distribution, in accord with previous estimates obtained with unequal-variance SDT models. Finally, the ratings' variance ratios correlated with the variance ratios obtained from an unequal-variance SDT model that was fit to the ratings data.



Note that the ratings' variance ratio agreed with the SDT-derived ratios even though both procedures are based on different assumptions. The SDT model, for example, assumes from the start that the familiarity distributions are Gaussian an assumption not shared by the ratings procedure used by Mickes et al. (2007). The fact that the results from the two procedures converged supports the assumptions of unequal-variance SDT models. In short, unequal-variance SDT models are plausible and are likely to provide appropriate fits to our recognition data.

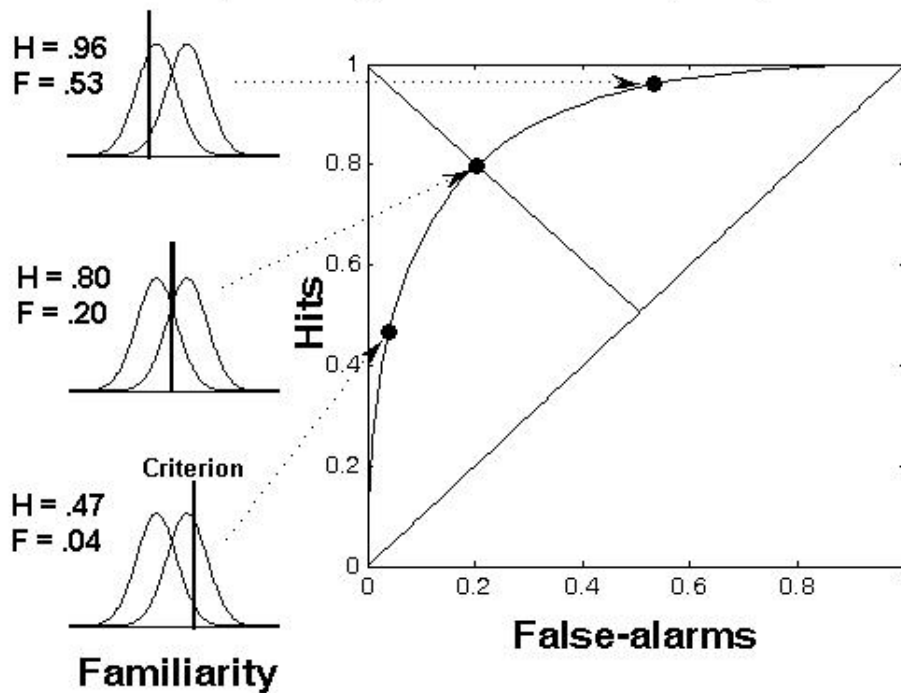
### 2.2.2. Receiver Operating Characteristic

Receiver Operating Characteristic (ROC) curves relate hit rates and false-alarm rates across changes in response criteria. They have been increasingly used in recognition memory because they provide, at a glance, a measure of sensitivity that is unaffected by the choice of a decision criterion (see Yonelinas & Parks, 2007, for a review of ROC studies in recognition).

An ROC can be constructed from an underlying SDT model by plotting hits against false alarms at progressively higher criterion locations along the familiarity continuum. Figure 2.2 illustrates the relationship between *target-lure* familiarity distributions and their corresponding ROC curve. The area under the ROC curve ( $A_z$ ) provides a measure of discriminability. When the ROC curve coincides with the main diagonal,  $A_z$  equals .5 (half the area of the unit square) and discriminability is at chance. When the ROC curve coincides with the left vertical and top horizontal axes,  $A_z$  equals 1.0 (the area of the unit square) and discriminability is perfect. Because  $A_z$  varies from .5 to 1, it can be interpreted as a proportion, adding an intuitive appeal that other measures, such as  $d_a$ , lack.

Theoretical ROC curves can be constructed simply by varying the criterion  $X$  from  $+\infty$  to  $-\infty$  along the familiarity continuum and plotting the corresponding hits and false alarms. To build empirical ROC curves, however, it is necessary to provide a method for experimentally manipulating criterion setting. *Old-new* and *confidence rating* tasks provide two different ways of manipulating the criterion.

## Receiver Operating Characteristic (ROC) Curve



**Figure 2.2. Equal variance SDT model and the ROC curve.**

An ROC can be constructed from an underlying equal-variance or unequal-variance SDT model by plotting hits against false alarms at progressively higher criterion locations along the familiarity continuum. Criterion location can be manipulated by rewarding participants differently for hits and false or by varying the proportion of *old* items on the test list. The area under the ROC curve ( $A_z$ ) provides a measure of discriminability. When the ROC curve coincides with the main diagonal,  $A_z$  equals .5 and discriminability is at chance; when the curve touches the top left corner of the scale,  $A_z$  equals 1.0 and discriminability is perfect.

### Old-new ROC

In old-new experiments, participants are presented with a test probe and have to decide whether or not the item was previously seen on the study list. The decision is binary (“old” vs. “new”) and is assumed to be based on both the level of familiarity elicited by the probe and the position of the criterion along the familiarity dimension. Participants tend not to change the criterion position during a recognition test, as attested by constant false-alarm rates across item-strength manipulations (e.g., Stretch & Wixted, 1998b). However, criterion setting can change dynamically during the course of a recognition test if participants are given trial-to-trial feedback on their performance (Rhodes & Jacoby, 2007; Verde & Rotello, 2007) or if the nature of the lure items changes dramatically and permanently during a testing sequence (Benjamin & Bawa, 2004; S. Brown, Steyvers, & Hemmer, 2007).

In most old-new recognition studies, criterion location has been manipulated by varying pay-offs or by varying the proportion of *old* items on the test list. Participants produce fewer “old” responses when higher penalties are applied on false alarms or when there is a high proportion of true *new* items on the test list. The problem with old-new experiments is that they require the collection of large amounts of data. That is because, in order to construct an ROC, hits and false alarms have to be obtained at each one of the manipulated confidence levels. Because of that, we use the alternative approach of collecting confidence ratings, which is less time-consuming and has been increasingly adopted over the last two decades (Yonelinas & Parks, 2007).

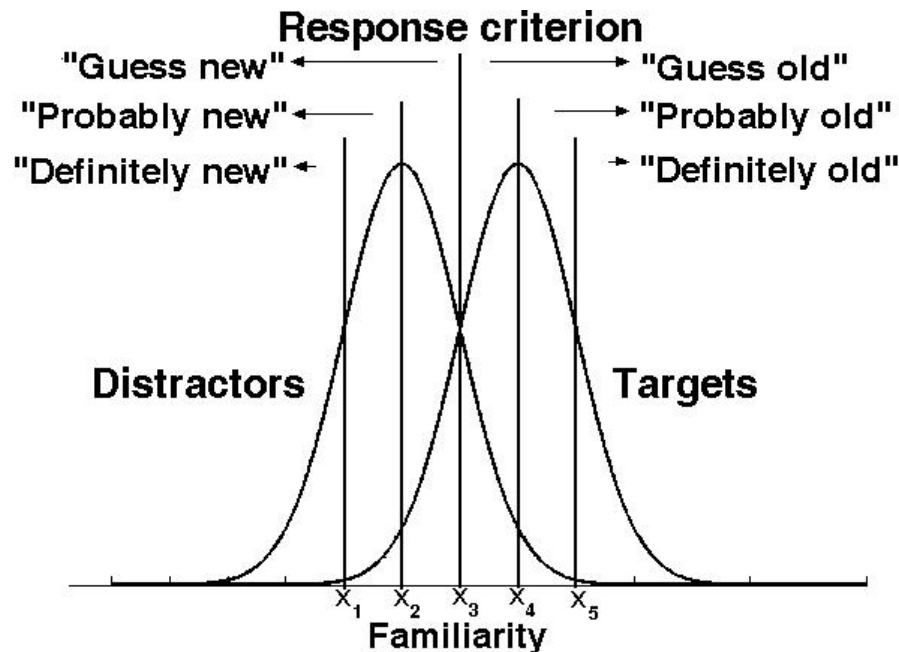
### Confidence rating ROC

In confidence ratings tasks, participants are asked to provide an index of perceived memory strength according to a fixed number of response categories. It is standard practice to use 6 categories labelled *definitely old*, *probably old*, *guess old*, *guess new*, *probably new* and *definitely new*. Entries to the first three categories represent “old” responses and entries to the last three represent “new” responses. Thus, confidence ratings provide information about subjective levels of memory strength in addition to old-new judgements.

If participants keep the boundaries between categories more or less constant throughout the recognition test (i.e., if they are consistent in the amount of familiarity they require in order to choose a particular category), then the category boundaries can be interpreted as response criteria. And if participants simultaneously keep those several criteria along the familiarity scale, then it becomes possible to measure responses based on different criteria in the same experiment, considerably reducing the amount of data necessary to produce an ROC curve. Figure 2.3 depicts an SDT model with 5 simultaneous criteria.

A confidence-rating ROC curve can then be produced from the SDT model in Figure 2.3. The model has 6 parameters (assuming  $\mu_D = 0$  and  $\sigma_D = \sigma_T = 1$ ): (i)  $\mu_T$ , the mean of the *target* distribution and (ii) the 5 criteria ( $X_1, X_2, X_3, X_4, X_5$ ) that separate “old” from “new” responses at each confidence level. The ROC is then produced by plotting hits and false alarms cumulatively across criteria. For

example, the leftmost point in the ROC is obtained by taking as hits the area to the right of the highest criterion ( $X_5$ ) in the *target* distribution and as false alarms the area to the right of that same criterion in the *distractor* distribution. The next ROC point is obtained by adding to the previous hit and false-alarm rates the area between the second highest criterion ( $X_4$ ) and the highest criterion ( $X_5$ ). Repeating this procedure for each criterion yields a concave<sup>1</sup> ROC similar to the one in Figure 2.2.



**Figure 2.3. Equal variance SDT model with five response criteria.**

In confidence rating experiments, it is assumed that participants place different response criteria simultaneously along the familiarity continuum and produce a rating (e.g., “probably old”) when the subjective level of familiarity elicited by a test probe falls within the corresponding region of the familiarity continuum (e.g., between criteria  $X_4$  and  $X_5$ ). Cumulative values of hits and false alarms at each confidence level are then used to construct theoretical ROC curves.

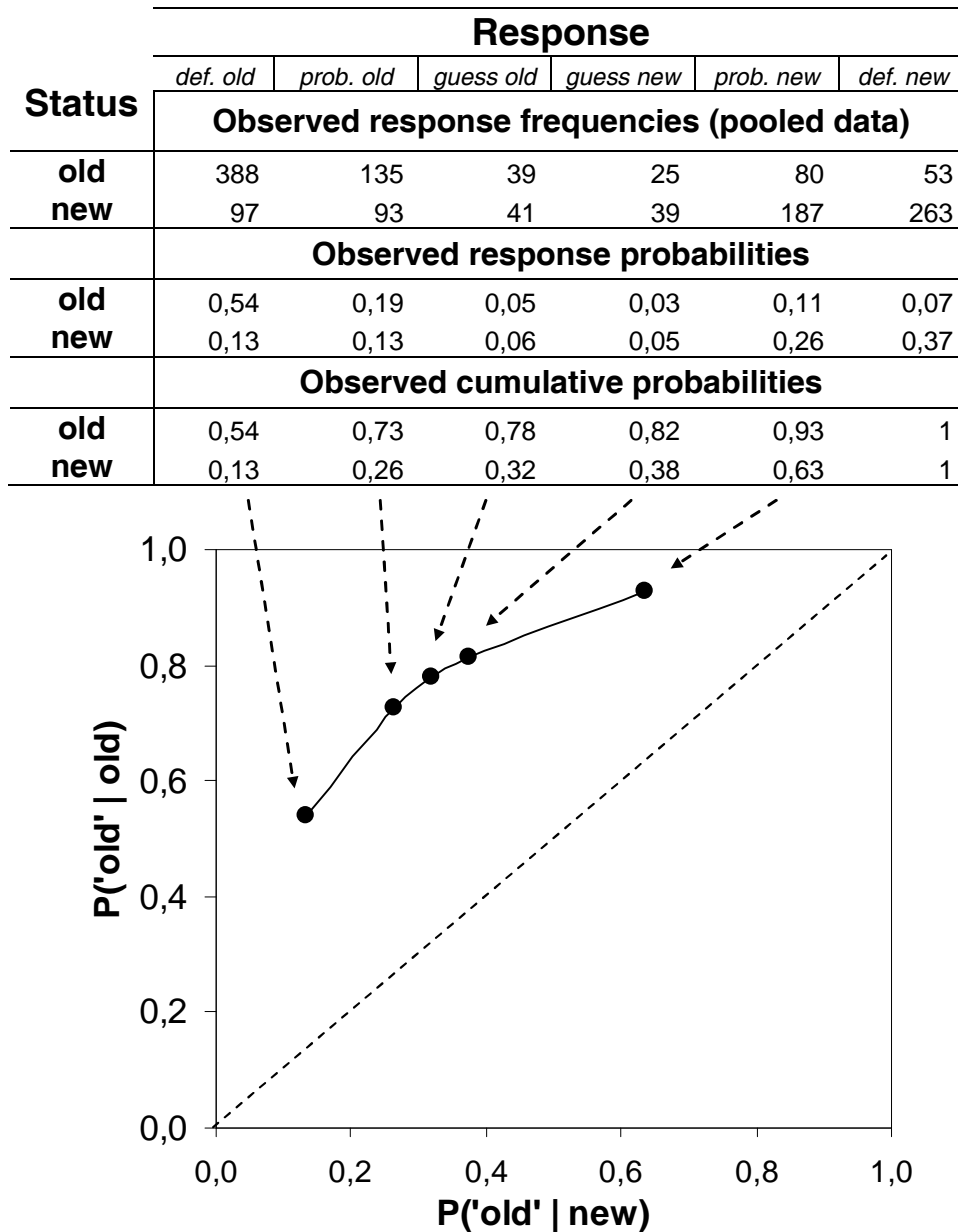
When the SDT model is not known in advance, which is the case in most recognition experiments, an *empirical ROC* can be constructed from participants’ data. Probabilities at each confidence level are estimated from the number of responses participants provide for each category (e.g., *probably old*) conditional on whether the item is a *target* or a *distractor*. The probabilities are then cumulatively added as if the boundaries between adjacent categories were SDT decision criteria. Figure 2.4 illustrates this procedure. The cumulative probabilities add to 1, so that the value at the rightmost category (e.g., *definitely*

<sup>1</sup> A function is concave if the y-value at the midpoint of the line segment connecting any two points on the function is less than the corresponding y-value on the function.

*new*) is uninformative. Thus, an ROC constructed from a 6-category confidence scale has in practice only 5 ( $F, H$ ) pairs (i.e., 5 degrees of freedom).

ROC curves constructed with confidence-based data tend to agree with curves constructed with old-new data (e.g., Ratcliff et al., 1992). However, the same manipulation may yield different results depending on whether the response requires a old-new or a confidence rating decision. Malmberg and Xu (2007), for example, showed in an associative recognition task that the number of false alarms to rearranged pairs increased monotonically with repetitions of the corresponding intact pairs when the task involved a old-new response; the number of false alarms, however, remained steady when the task involved a confidence rating task. They hypothesised that the faster responses obtained in old-new tasks indicated lower engagement of recollection, resulting in poorer performance. Malmberg and Xu (2007) tested this idea by forcing participants in the old-new task to wait 2 s before entering a response and found that false alarms indeed increased less with intact-pair repetition, approaching the pattern of responses obtained with confidence ratings. Thus, old-new and confidence ratings tasks may alter the relative engagement of memory processes (in this case, recollection), possibly yielding different results.

Another reason to be careful with confidence-based data is that they entail the assumption that participants can map subjective levels of confidence directly to levels of memory strength. This assumption, however, has been shown to be incorrect in general. Van Zandt (2000) showed that measures of variance and sensitivity ( $z$ ROC slope and intercept, see below) changed when she manipulated response criteria through pay-offs and percentage of *old* items at test. In particular, the variance of the *target* distribution increased and sensitivity decreased in the more conservative condition (i.e., where participants were discouraged to say “old”). The result contradicts the SDT prediction that familiarity distributions should not be affected by changes in criterion placement.



**Figure 2.4. Construction of ROC curve from frequency data.**

Receiver Operating Characteristic (ROC) curves can be constructed from participants' data by plotting hits [ $P(\text{"old"}|\text{old})$ ] against false alarms [ $P(\text{"old"}|\text{new})$ ] across different confidence levels. Probabilities at each confidence level are estimated from the number of responses in each category (e.g., *probably old*) conditional on whether the item is *old* or *new*. The probabilities are then cumulatively added as if the boundaries between adjacent categories were SDT decision criteria. The ROC data points are then used by RscorePlus to find best-fitting SDT parameters.

More important to our purposes, the drop in sensitivity with more conservative responding is a potential confound in list-strength studies because list strength normally causes participants to become more conservative (Hirshman, 1995). Taken together, Van Zandt's (2000) and Hirshman's (1995) results suggest that LSEs may occur because of distortions in the mapping between confidence and memory strength in *mixed* lists compared to *pure weak* lists, not because strong

items interfere with weak items. However, in the experiments reported in Chapters 3 and 4, no systematic differences in estimated variance were observed across list types, suggesting that, although list strength can change criterion setting between lists, it may not be as harmful to the confidence-to-familiarity mapping as other criterion manipulations (see Verde & Rotello, 2007, for evidence that strength alone is not sufficient to cause within-list criterion shifts).

### Normalised ROC (zROC)

Although ROC curves provide useful information about discriminability, they are ill-suited to describe the variances of the underlying distributions. When the variances of *target* and *lure* distributions are different, ROC curves become asymmetric with respect to the minor diagonal. Quantifying the asymmetry of ROC curves, however, is not straightforward. The conversion of ROC curves into zROC curves allows the measurement of that asymmetry and thus of the relative variances of the underlying distributions, in a relatively easy manner.

zROC curves are ROC curves in which hits and false alarms were transformed by the inverse cumulative Gaussian distribution operator  $z(P)$ , where  $P$  is a probability value (e.g., hit rate  $H$  or false-alarm rate  $F$ ).  $z(P)$  returns the point  $z$  on the familiarity scale for which  $P = \Phi(z)$ . The function  $\Phi(z)$  is the cumulative Gaussian distribution and corresponds to the area under the distribution to the left of cut-off point  $z$ . This area, which represents a probability, is given by:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}} dx \quad (2.1)$$

False alarms are related to the cumulative Gaussian distribution by the expression  $F = 1 - \Phi((X_i - \mu_D)/\sigma_D)$ , which is the area of the Gaussian function to the *right* of criterion  $X_i$  with respect to a *lure* distribution with mean  $\mu_D$  and standard deviation  $\sigma_D$ . Hits are related to the cumulative function by the expression  $H = 1 - \Phi((X_i - \mu_T)/\sigma_T)$ , which is the area to the *right* of criterion  $X_i$  with respect to the *target* distribution with mean  $\mu_T$  and standard deviation  $\sigma_T$ .

Because the Gaussian distribution is symmetric, it is possible to rewrite those expressions such that  $F = 1 - \Phi((X_i - \mu_D)/\sigma_D) = \Phi((\mu_D - X_i)/\sigma_D)$  and  $H = 1 - \Phi((X_i - \mu_T)/\sigma_T) = \Phi((\mu_T - X_i)/\sigma_T)$ .

$-\mu_T)/\sigma_T) = \Phi((\mu_T - X_i)/\sigma_T)$ .  $z$ ROC coordinates can then be related to the model parameters (means and standard deviations) by applying  $z(P)$ , so that  $z(F) = (\mu_D - X_i)/\sigma_D$  and  $z(H) = (\mu_T - X_i)/\sigma_T$ . Finally, eliminating  $X_i$  from the equations yields

$$z(H) = \frac{\mu_T - \mu_D}{\sigma_T} + \frac{\sigma_D}{\sigma_T} z(F) \quad (2.2)$$

which is a straight line in  $z$ -space, if the underlying distributions are Gaussian<sup>2</sup>, with *intercept*  $(\mu_T - \mu_D)/\sigma_T$  and *slope*  $\sigma_D/\sigma_T$ . The  $z$ ROC slope provides a measure of the relative sizes of *target* and *lure* variances. When the slope is less than 1, *target* variance is greater than *lure* variance.  $z$ ROCs fitted to empirical ROCs have generally produced slopes revolving around 0.8 (e.g., Ratcliff et al., 1992). So, estimated *target* variance is normally 1.25 times greater than *lure* variance.  $z$ ROCs were produced from our data by fitting a straight line through  $z(H)$  and  $z(F)$  coordinates (see 2.3.2 for details).

### 2.2.3. Sensitivity measures ( $d'$ , $d_a$ , $A_z$ )

*Sensitivity* refers to the ability to *discriminate* between *targets* and *lures*.<sup>3</sup> When *lure* and *target* distributions have the same variance ( $\sigma_D = \sigma_T$ ), sensitivity can be measured as the standardised distance between the distribution means, given by  $(\mu_T - \mu_D)/\sigma_T$ , which is the  $z$ ROC intercept in Equation 2.2. This measure, called  $d'$  in Equation 1.1, can also be expressed in terms of  $z$ ROC coordinates by rearranging Equation 2.2 and assuming that  $\sigma_D = \sigma_T$ . The resulting expression

$$d' = z(H) - z(F) \quad (2.3)$$

is commonly used in the recognition memory literature, partly because it is easy to compute (e.g., with Excel) and partly because it requires only one set of hits and false alarms (i.e., a single response criterion). The use of  $d'$ , however, entails the assumption of equal variance, which is not empirically valid (see 2.2.1).

More importantly, when variances are unequal,  $d'$  varies with criterion placement, which opens up the possibility that the sensitivity measure may change between conditions known to affect bias (e.g. *short* vs. *strong* lists),

<sup>2</sup> Non-Gaussian functions also produce linear  $z$ ROCs (Lockhart & Murdock, 1970). Van Zandt (2000) showed that exponentially distributed *targets* and *lures* can yield linear  $z$ ROCs.

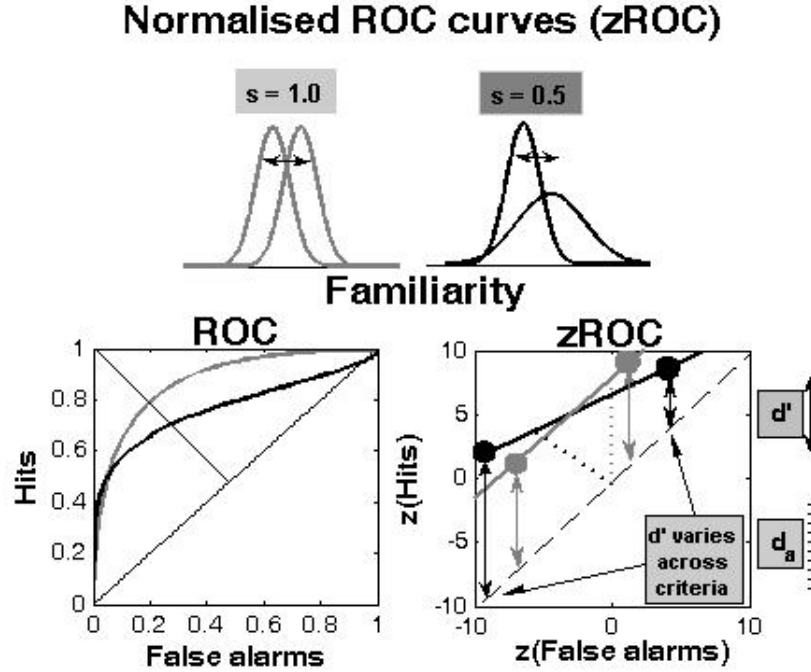
<sup>3</sup> The terms sensitivity and discriminability are used interchangeably in this thesis.



despite no real change in recognition sensitivity. Indeed, Verde and Rotello (2003) found that to be the case when investigating the *revelation effect*, the finding that participants respond “old” more often after doing an unrelated task prior to the recognition test. Although previous studies had shown changes in  $d'$  between revelation (with task) and no revelation (without task) conditions, suggesting a change in familiarity, Verde and Rotello (2003) showed that the incidental task was affecting criterion placement rather than sensitivity and that previous data reporting changes in  $d'$  were likely to be a by-product of shifts in bias, without any real changes in familiarity. Figure 2.5 illustrates this point. Owing to this type of misbehaviour, single-point sensitivity measures such as  $d'$  are not appropriate for our purposes. Nonetheless,  $d'$  data are reported in Appendix 1 for completeness and to allow comparisons with previous studies.

What is needed then is a measure that takes into account the variances of both *target* and *lure* distributions and that is robust against criterion shifts. One such measure is called  $d_a$ . Sensitivity  $d_a$  takes into account the two variances by using an average variance (*root mean square*:  $\sqrt{(\sigma_D^2 + \sigma_T^2)/2}$ ), rendering  $d_a$  more robust to criterion shifts than  $d'$  (see the dotted lines on the  $z$ ROC in Figure 2.5).

Sensitivity  $d_a$  can be calculated with analytical geometry (Wickens, 2002, p. 65). The goal is to measure the distance between a straight line (the  $z$ ROC curve) and the *indifference curve* (main diagonal in  $z$ -space, where discriminability is zero). The farther the  $z$ ROC curve is from the diagonal, the higher is the sensitivity. The problem is that the distance between the lines is not constant when the  $z$ ROC slope is different from 1. One approach is to determine the *minimum distance* between the  $z$ ROC curve and the origin point (0,0) in  $z$ -space and to *scale* that distance to reflect a compromise between the two underlying variances.



**Figure 2.5. Effect of criterion shifts on single-point sensitivity.**

Sensitivity  $d'$  should be used only when the variance of the underlying distributions are equal. When the variances are unequal, sensitivity  $d'$  can change depending on criterion location. In the example, when the variance ratio  $s (= \sigma_D / \sigma_T)$  is 1.0 (equal variances),  $d'$  remains unchanged, regardless of criterion position [i.e., distance (double arrows), on  $z$ -space, between the equal-variance line (gray) and the main diagonal is constant]; when the variance ratio  $s$  is 0.5 (target variance twice as large as distractor variance),  $d'$  changes depending on criterion position [i.e., distance between the unequal-variance line (black) and the diagonal is not constant]. Sensitivity  $d_a$ , which measures the distance between *target* and *lure* distribution means in units of the *root mean square* of their standard deviations, is more resilient to shifts in criterion placement.

The minimum distance ( $d_{min}$ ) between a line ( $zROC$ ) and a point (origin) lies in the direction perpendicular to the line. Let  $a = (\mu_T - \mu_D) / \sigma_T$  and  $b = \sigma_D / \sigma_T$  [so that Equation 2.2 can be rewritten as (i)  $z(H) = a + b z(F)$ ]. It is known from analytical geometry that a line perpendicular to another line with slope  $b$  has a slope of  $-1/b$ ; moreover, this perpendicular line passes through (0,0), so its intercept is 0. Thus, the perpendicular line is defined by (ii)  $z(H) = -(1/b) z(F)$ . The point at which the  $zROC$  line and its perpendicular counterpart intersect is obtained by solving (i) and (ii). The distance between the origin (0,0) and the calculated intersection point  $A = [-ab/(1+b^2), a/(1+b^2)]$  is (iii)  $a/\sqrt{1+b^2}$ . Replacing  $a$  and  $b$  in (iii) yields  $d_{min} = (\mu_T - \mu_D) / \sigma_T \sqrt{1 + \sigma_D^2 / \sigma_T^2}$ , the minimum distance between the  $zROC$  and the origin. If the variances are equal, however,

$d_{min}$  becomes  $(\mu_T - \mu_D)/\sigma_T\sqrt{2}$ , which is  $d'/\sqrt{2}$ . Thus, to produce a value of the magnitude of  $d'$ ,  $d_{min}$  needs to be rescaled. Multiplying  $d_{min}$  by  $\sqrt{2}$  results in

$$d_a = \sqrt{2} d_{min} = \sqrt{2} \frac{(\mu_T - \mu_D)}{\sigma_T\sqrt{1 + \sigma_D^2/\sigma_T^2}} = \frac{(\mu_T - \mu_D)}{\sqrt{(\sigma_D^2 + \sigma_T^2)/2}} \quad (2.4)$$

which represents the distance between *target* and *lure* means measured in units of the *root mean square* of their standard deviations. When variances are the same, Equation 2.4 reduces to Equation 1.1; when variances are different, the measure takes into account their relative values by averaging them.  $d_a$  can also be obtained directly from a point on the ROC once the zROC slope  $s$  is known ( $s = \sigma_D/\sigma_T$ ). If the intercept  $(\mu_T - \mu_D)/\sigma_T$  is eliminated from Equations 2.2 and 2.4, then  $d_a = \sqrt{(2/(1 + s^2))} [z(H) - s z(F)]$ . Because  $d_a$  varies less than  $d'$  with criterion shifts,  $d_a$  is a more suitable measure for us.

In practice, sensitivity will be reported in this thesis using  $A_z$ , a measure closely related to  $d_a$ . Recent work confirmed that  $A_z$  possesses better statistical properties than  $d'$  (Verde, Macmillan, & Rotello, 2006).  $A_z$  not only captures the advantages of  $d_a$  in relation to  $d'$  but it also provides a measure that can be interpreted as a proportion.  $A_z$  represents the area under the ROC curve and varies from .5 (chance performance; area of main diagonal) to 1 (perfect performance, area of unit square).  $A_z$  assumes Gaussian distributions and is obtained from  $d_a$  by

$$A_z = \Phi(d_{min}) = \Phi\left(\frac{d_a}{\sqrt{2}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{d_a}{\sqrt{2}}} e^{-\frac{x^2}{2}} dx \quad (2.5)$$

#### 2.2.4. Bias measures ( $X_i$ , $c$ , $c_a$ )

Response bias is defined here as the tendency of the participant to respond “old”. Consequently, a measure of bias should incorporate information about both hits (say “old” to *old* items) and false alarms (say “old” to *new* items). Moreover, this bias measure should co-vary with hits and false alarms, such that it rises if those raw measures rise and it falls if the raw measures fall. Thus, unlike sensitivity, which depends on the *difference* between hits and false alarms, bias should depend on the *sum* of those two measures.

The simplest measure of bias involves taking only false alarms into account.

Recall that  $z(F) = (\mu_D - X_i)/\sigma_D$ . So, criterion  $X_i$  can provide a measure of bias given by  $X_i = \mu_D - z(F)\sigma_D$ . If we assume  $\mu_D = 0$  and  $\sigma_D = 1$ , then

$$X_i = -z(F) \quad (2.6)$$

Thus, higher false-alarm rates indicate lower response bias and lower false-alarm rates indicate higher response bias. Participants are called *liberal* in the first case and *conservative* in the second case. The problem with this measure is that it does not take into account information about the *target* distribution. Not only is this bias measure insensitive to hits but it also does not reflect any strategy a participant may use to maximise performance.

For example, a conservative strategy (high  $X_i$ ) may be good if discrimination is high, as a few misses would be compensated by fewer false alarms. The same conservative strategy, however, may not be so good if discrimination is low, as the large proportion of misses would overshadow the benefits of few false alarms. In fact, participants appear to tune in to information from both *target* and *lure* distributions in order to set their response criteria (Curran, DeBuse, & Leynes, 2007; Rhodes & Jacoby, 2007). Thus, a measure of bias should also take into account hit rate information.

Hits and false alarms can be combined into a single measure by including  $z(F)$  and  $z(H)$  in the same expression. However, it is important to determine first the point of *zero bias* relative to which participants' responses can be classified as liberal or conservative. One possibility is to define the point of zero bias as the point that maximises the number of correct responses. The probability  $P_C$  of a correct response is the probability of saying “old” in a *target* trial and “new” in a *lure* trial. Thus,  $P_C = P(\text{target})P(\text{“old”} | \text{target}) + P(\text{lure})P(\text{“new”} | \text{lure})$ . Because  $P(\text{target}) + P(\text{lure}) = 1$  and given that, by definition,  $H = P(\text{“old”} | \text{target})$  and  $1 - F = P(\text{“new”} | \text{lure})$ , it follows that  $P_C = P(\text{target})H + [1 - P(\text{target})](1 - F)$ . Recall that  $H = 1 - \Phi((X_i - \mu_T)/\sigma_T)$  and  $1 - F = \Phi((X_i - \mu_D)/\sigma_D)$ . It can be shown that, when  $\mu_D = 0$ ,  $\sigma_D = \sigma_T = 1$  and  $P(\text{target}) = .5$  (i.e., half of the test items are *targets*), the criterion  $X$  that maximises  $P_C$  is such that it yields  $H = 1 - F$ , which is the midpoint between *lure* and *target* distributions in the equal-variance model.

At that point, where hits equal misses, bias is set to zero. For the equal-variance SDT model, a bias measure  $c$  can then be derived such that

$$c = -\frac{1}{2}[z(H) + z(F)] \quad (2.7)$$

Bias  $c$  has some of the desired properties. It depends on the sum of hits and false alarms, is negative when responding is liberal and positive when responding is conservative, and is zero when the condition for maximum correct responses is met [ $c = 0$  when  $H = 1 - F$  because  $z(H) = z(1 - F) = -z(F)$ ].

When variances are different, however, a more appropriate measure is given by

$$c_a = -\frac{\sigma_T}{\sigma_D + \sigma_T} \sqrt{\frac{2}{\sigma_D^2 + \sigma_T^2}} [z(H) + z(F)] \quad (2.8)$$

which uses as an average variance the root mean square of *distractor* and *target* variances.  $c_a$  reduces to  $c$  when  $\sigma_D = \sigma_T = 1$ . When hits and false alarms are plotted in an ROC, each  $(F_i, H_i)$  pair defines a corresponding bias  $c_a(F_i, H_i)$ . For each model fitted to our data, there are five criteria  $X_i$  along the familiarity scale generating 5  $(F_i, H_i)$  pairs and 5  $c_a(F_i, H_i)$  values. In this thesis, whenever we refer to bias  $c_a$ , we are in fact referring to the bias associated with criterion  $X_3$ , which is the criterion in the middle and should represent a better estimate than  $c$  when variances are different. Results for bias  $c$  are provided in Appendix 1.

## 2.3. Data analysis

### 2.3.1. Raw measures

Hits, false alarms and response times provide the most direct measures of performance in our experiments. These raw measures, together with the derived measures  $A_z$  and  $c_a$ , are analysed using standard statistical techniques, such as Analysis of Variance (ANOVA) and  $t$ -tests (Howell, 2002).

Analysing hits and false alarms separately is relatively assumption-free (apart from the assumptions of the statistical test being used). Combining hits and false alarms, however, entails some assumptions that deserve mention. Hits and false alarms are combined when they are entered in an ANOVA carried out on the proportion of “old” responses having *word type* (*target* vs. *lure*) as an

independent variable. Suppose the goal is to determine whether list length affects hits and false alarms. By entering *length* (*short* vs. *long*) as a second independent variable in the ANOVA, it is possible to assess whether the variables *word type* and *list length* interact such that the proportion of hits (“old” responses to targets) is lower in the *long* list than in the *short* list and the proportion of false alarms (“old” responses to lures) is higher in the *long* list than in the *short* list. Another way of looking at this interaction, however, is to ask whether the difference between hits ( $H$ ) and false alarms ( $F$ ) is lower in *long* lists than in *short* lists. In other words, the interaction is assessing whether sensitivity, measured as  $H - F$ , differs across lists. The question that arises is how good  $H - F$  is as a measure of sensitivity and what assumptions its use implies. The short answer is that  $H - F$  is not a good measure of sensitivity and that its assumptions are not warranted.

Suppose that after a study item is presented, a participant’s memory of the item exists in only one of three mutually exclusive states: ( $O$ ) the item is stored; ( $N$ ) the item is not stored; ( $U$ ) the status of the item is uncertain. Suppose further that there is a threshold that must be reached in order for an *old* item to be detected as old and another threshold for a *new* item to be detected as new. Finally, assume that only *old* items are able to reach state  $O$  (with probability  $P_O$ ) and that only *new* items are able to reach state  $N$  (with probability  $P_N$ ). When *targets* at test do not reach state  $O$ , memory falls into state  $U$ . Likewise, when *lures* do not reach state  $N$ , memory also falls into state  $U$ . For simplicity, let  $P_O = P_N = P_d$ .

Probability  $P_d$  is a measure of sensitivity as it represents the probability of correctly accepting *targets* and rejecting *lures*. Hits can be produced either when a *target* leads to state  $O$  or when a *target* leads to state  $U$  and an “old” response is guessed. If  $P_U$  is the probability of guessing “old”, then the probability of a hit is given by  $H = P_d + (1 - P_d)P_U$ . False alarms can be generated only by a *lure* that fails to reach state  $N$  and falls into state  $U$ ; that is, an “old” response to a *lure* is generated only by guesses, such that  $F = (1 - P_d)P_U$ . Note that the ROC for this threshold model is given by a straight line ( $H = P_d + F$ ). Isolating sensitivity  $P_d$  yields  $P_d = H - F$ . Thus, the use of  $H - F$  as a sensitivity measure implies the assumption that the familiarity signal is generated by a threshold model.

Threshold models differ from signal-detection models in that they assume only a finite number of states of memory (as opposed to the infinite number of memory states allowed in the familiarity continuum) and that criterion placement simply reflects the probability of guessing “old” in the absence of *any* other information about the test item (as opposed to reflecting the control over the proportion of “old” responses based on *some* information about the item). Previous research (e.g., Macho, 2004) and our own results show that threshold models such as the two-high threshold model above provide poor fits to the data.

This long discussion serves the purpose of justifying our preference for derived measures when drawing conclusions about the effects of interest. Although we do report *word type*  $\times$  *list type* interactions in all our experiments, we do not make strong claims about them, especially about comparisons between *short* and *strong* lists, as they are likely to confound memory factors with decision factors. Thus, for the purposes of this thesis, list-length (LLE) and list-strength effects (LSE) are defined in terms of sensitivity  $A_z$ . By contrast, significant interactions between word type [*target* (*H*) vs. *lure* (*FA*)] and list type (*short*, *long*, *strong*) on the proportion of “old” responses will be simply referred to as an effect of length or strength manipulations, not as LLE or LSE.

### 2.3.2. Derived measures

In this section we describe how the derived measures of sensitivity ( $A_z$ ) and bias ( $c_d$ ) were estimated from raw data. It is important to note that the signal-detection estimates  $A_z$  and  $c_d$  were obtained from fits to *individual participants’ data*, meaning that an ROC was produced for each participant. The average estimates across participants were then used in the statistical analyses. This procedure contrasts with analyses of *aggregate data*. An aggregate ROC is constructed with the data from all participants lumped together (see Figure 2.4 for an example). It is known, however, that aggregating data in this way may distort model estimates (S. Brown & Heathcote, 2003; Estes & Maddox, 2005; Malmberg & Xu, 2006). The aggregate ROCs reported in Chapters 3 and 4 are provided for illustrative purposes only. None of the conclusions in this thesis depends on the aggregate

data. The measures derived from the aggregate ROCs, however, did not substantially differ from the measures derived from individual participants' ROCs. This apparent lack of averaging artifacts not only is reassuring but also provides evidence that some of the necessary corrections inflicted on individual participants' raw data (e.g., replacing extreme values of hits and false alarms or replacing zero entries in some response categories) did not substantially alter the pattern of results.

### Parameter estimation (RscorePlus)

The confidence ratings collected at test were used to construct ROC curves for each participant and list type. Sensitivity ( $A_z$ ) was estimated by fitting an unequal-variance Gaussian model to each participant's confidence data. Unequal-variance models tend to provide good fits to ROC data (Wixted, 2007). The best fitting model was obtained with the RscorePlus maximum-likelihood algorithm (Dorfman & Alf, 1969; Harvey, 2001, <http://psych.colorado.edu/~lharvey>).

The RscorePlus algorithm assumes  $\mu_d = 0$  (mean familiarity of *distractors*) and  $\sigma_d = 1$  (standard deviation of *distractors*). It also assumes that the observer holds 5 decision criteria along the familiarity space. The algorithm then estimates  $\mu_t$  (mean familiarity of *targets*),  $\sigma_t$  (standard deviation of *targets*) and the 5 criteria ( $X_i$ ;  $i = 1, \dots, 5$ ) and associated biases ( $c_a$ ) relative to the *distractor* distribution. The set of decision criteria and biases were estimated from the participant's entries on the 6-point rating scale at test.<sup>4</sup> The bias results in this thesis refer to the measure associated with the third criterion ( $X_3$ ) estimated for each participant.

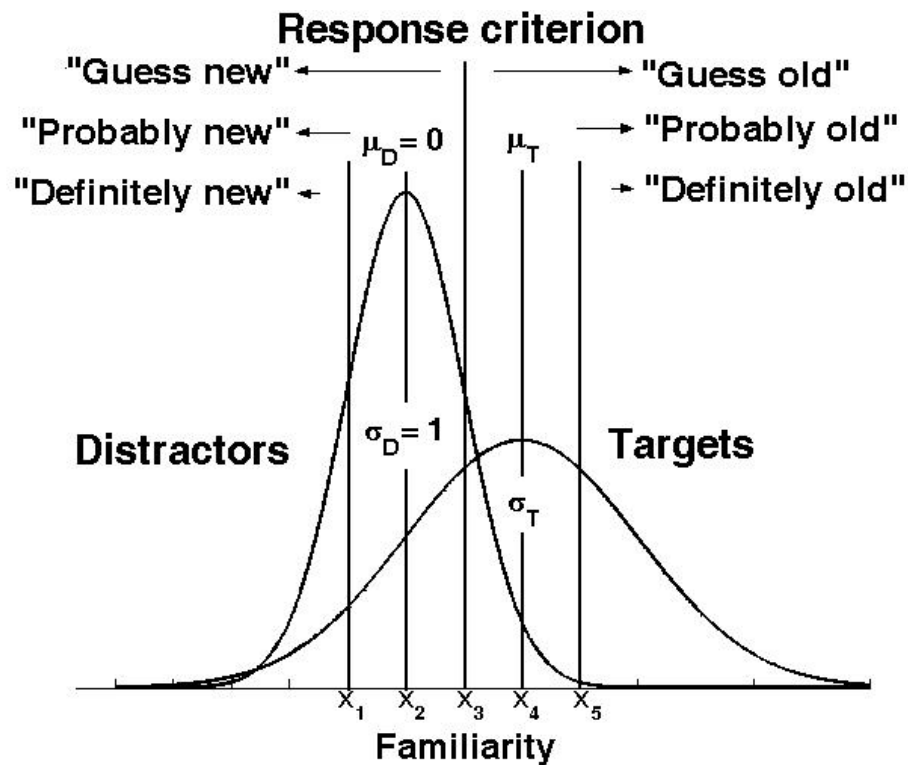
$z$ ROC slopes ( $\sigma_n/\sigma_o$ ), which represent the ratio between *new* and *old* distribution variances, were estimated by fitting a straight line through  $z(H)$  and  $z(F)$  coordinates for each model. Strictly speaking, standard linear regression is not the most appropriate method to fit ROC data because it assumes that data vary only in the y-dimension and that the each data point is independent, whereas ROC data vary in both x- and y-dimensions and ROC data points are not independent, since they are obtained by cumulatively adding hits and false

---

<sup>4</sup> If participants failed to enter responses to a particular confidence rating (e.g., *guess old*), the zero frequency value was replaced with 0.17 (= 1/6) to avoid collapsing data across rating values.



alarms across decision criteria. Despite these caveats, linear regression is used here due to its simplicity and due to the fact that linear regression fits are very similar to the more appropriate maximum-likelihood fits (Ratcliff et al., 1994). For undefined values in  $z$ -space, a standard correction was applied (Macmillan & Creelman, 2005, p. 21): if  $H = 1$ , then  $H_{corr} = 1 - 1/2n$ , where  $n$  is the number of trials per word type; if  $F = 0$ , then  $F_{corr} = 1/2n$ . Similarly, if  $H = 0$ , then  $H_{corr} = 1/2n$  and if  $F = 1$ , then  $F_{corr} = 1 - 1/2n$ .



**Figure 2.6. Unequal-variance SDT model estimated by RscorePlus.**

The RscorePlus maximum-likelihood algorithm takes as input the confidence ratings produced by participants at test. The algorithm assumes that the mean and standard deviation of the *distractor* distribution are fixed ( $\mu_D = 0$ ,  $\sigma_D = 1$ ). Then, it searches for the values of the mean and standard deviation of the *target* distribution ( $\mu_T$ ,  $\sigma_T$ ) and for the 5 criteria ( $X_i$ ,  $i = 1$  to 5) that maximise the likelihood of the data provided by participants. Because ratings are cumulatively added in the ROC, only 5 out of the 6 ratings are free to vary. Thus, for each of RscorePlus fit, there are 10 data points to fit [5 ratings  $\times$  2 word types (*target*, *distractor*)] and 7 parameters to estimate.

To summarise, the derived measures were computed as follows. First, ROC curves were generated for each participant. Second, an unequal-variance Gaussian model was fitted to each ROC curve. Finally, the model parameters ( $\mu_T$ ,  $\sigma_T$  and  $X_i$ ) were used to calculate  $A_z$ , and  $c_a$ . Participants whose data provided poor fits in *at least* one condition were filtered out (chi-squared  $p$ -value  $< .05$ ; note that this is a very liberal exclusion criterion). The pattern of results was not

altered by including the data from poor fits in the analysis (although they may have changed the significance values); for brevity, we do not report these results. Alpha was set to .05 (two-tailed) for all analyses, unless otherwise stated.

### Comparisons of interest

The experiments in Chapters 3 and 4 seek to assess whether or not there are differences in sensitivity  $A_z$  between *short* lists (baseline condition; items presented once), *long* lists (longer than *short* lists; items presented once) and *strong* lists (same size as *long* lists but with as many unique items as *short* lists; half the items presented more than once).

Testing the hypotheses outlined in the Introduction requires three comparisons: (1) the comparison between *studied items* and *unrelated lures* (SU), (2) the comparison between *studied items* and *switched-plurality (SP) lures* (SSP), and (3) the comparison between *SP lures* and *unrelated lures* (SPU). Although these comparisons do not necessarily entail pure processes, they are still thought to involve different contributions of *familiarity* and *recollection*. Consequently, each comparison yields a relative measure of potential differential effects of list-length and list-strength manipulations on familiarity and recollection.

### **SU comparison – Studied vs. Unrelated lures**

The SU comparison provides a measure of how likely are participants to say “old” to *targets* compared to *unrelated lures*. Because the distinction between *targets* and *unrelated lures* can be made with familiarity alone, this comparison provides a rough index of familiarity-dependent discrimination for each list type.

Note that recollection can also play a role in this type of discrimination. For example, participants may see the lure *coat*, recall that they saw the similarly sounding item *boat* and use that information to infer whether or not *coat* was studied. Participants may use this recalled information in at least two ways: *i*) if they believe that similarly sounding words were infrequent, they can use the recalled information to reject the test item as “new”; *ii*) if the level of subjective memory strength elicited by recalling *coat* is used as an anchor to decide whether or not the strength elicited by *boat* is high enough to warrant an “old” response.

However, due to the nature of the study lists used here, which were composed of randomly similar items, the proportion of trials in which such *diagnostic* (or heuristic) use of recall at test should occur is deemed to be small (see Gallo, 2004, for a distinction between diagnostic and disqualifying uses of recall).

### **SSP comparison – Studied vs. Switched-Plurality lures**

The SSP comparison provides a measure of how likely are participants to say “old” to *targets* compared to *SP lures*. Because the correct distinction between *targets* and *SP lures* may require recollection of plurality information (in addition to familiarity), this comparison provides a rough measure of recollection-dependent discrimination for each list type.<sup>5</sup>

Evidence that plurality discrimination involves a recollective component comes from response-signal studies which showed that it takes longer to reject *SP lures* than *unrelated lures* (Hintzman & Curran, 1994), a dynamic that is also found in associative recognition (Gronlund & Ratcliff, 1989), which behaves like recall in many tasks. The same temporal dynamics has been seen in electrophysiological studies which, in addition, found evidence pointing to different neural processes underlying *unrelated* and *SP lure* discrimination (Curran, 2000). The early event-related component (300 – 500 ms; FN400) showed no difference between “old” responses to *targets* and *SP lures*, suggesting that both word types elicited similar levels of familiarity. Moreover, the FN400 component was higher for hits and SP false alarms than for unrelated false alarms, suggesting that familiarity for *unrelated lures* was lower than for *targets* and *SP lures*. By contrast, the late event-related component (400 – 800 ms; parietal) showed that “old” responses to *targets* elicited higher signal amplitudes than “old” responses to *SP lures* and *unrelated lures* (see also Curran & Cleary, 2003), suggesting that it mediates the correct recall of studied items. Differential activation of the late parietal component has also been observed in deep-encoding tasks (e.g., sentence generation), which are known to elicit high levels of recollection, but not in

---

<sup>5</sup> It is possible to describe SSP discrimination with familiarity only. Heathcote et al. (2006) proposed a model in which it is assumed that memory is probed twice at each trial (once with a singular probe and once with a plural probe) and that the difference in those matches is used to construct a strength-of-evidence dimension, based on which recognition decisions are made.

shallow-encoding tasks (e.g., letter judgement), which elicits less recollection (see Rugg & Curran, 2007 for a review). Thus, time-course and physiological studies suggest that SSP discrimination can trigger recollection.

The role of recollection in *SP lure* discrimination is also supported by aging studies. Aging affects recollection more than familiarity (e.g., Prull et al., 2006); and a similar pattern is found for *SP lure* discrimination across age groups. In particular, aging appears to impair the ability to use recall-to-reject strategies against *SP lures*. When *targets* are presented repeated times at study, false alarms to corresponding *SP lures* remain steady (or decrease slightly) for young adults but increase for old adults (Light, Chung, Pendergrass, & Van Ocker, 2006). This pattern can be readily interpreted by assuming that young adults can successfully recruit recollection to counter the increased familiarity of repeated *targets*, whereas old adults are less successful in recruiting recollection. This interpretation is supported by the fact that some single-process models had to include a recall component in order to correctly model the steadiness of SP false alarms (e.g., REM model: Malmberg, Holden et al., 2004). In sum, there is strong evidence supporting the view that SSP discrimination involves recollection and that it does so to a higher degree than SU discrimination.

Unlike SU discrimination, where recalling an item may or may not help rejecting a lure, in SSP discrimination the recall of an item allows *targets* to be accepted and lures to be rejected; if the participant recalls *bananas* upon presentation of *banana*, he can be sure that the item is a lure. However, to use recall in this *disqualifying* (or rule-based) manner (Gallo, 2004), participants have to know about its disqualifying value. Thus, it is important to instruct participants to pay attention to plurality information at study. Moreover, it is crucial to mention that items at study are presented either in their singular or plural form, but not both, making it clear that any recollected information at test can be used to reject similar lures. Indeed, previous studies showed that the consistent engagement of recollection at test may be under strategic control (Gallo, 2004; Rotello et al., 2000; Westerman, 2000). In keeping with those results, we explicitly told our participants about *SP lures* and instructed them to use recall-to-reject at test.

Note that the same *targets* were used in both SU and SSP comparisons. The only difference between the comparisons lies in the nature of the lures (*unrelated* vs. *SP lures*). If list-length and list-strength manipulations act by specifically reducing recollection (assuming it is a recall-like process), we should observe a decrease in performance in SSP comparisons, where recollection is more likely to occur, but not in SU comparisons, where familiarity alone may be sufficient for correct old-new discrimination.

### **SPU comparison – Switched-Plurality lures vs. Unrelated lures**

In the SPU comparison, *SP lures* are analysed as *targets* (*pseudotargets*). This provides a measure of *pseudodiscrimination* (i.e., how more likely are participants to say “old” to *SP lures* compared to *unrelated lures*). High levels of recollection should produce low pseudodiscrimination because *SP lures* would be confidently rejected. In contrast, low levels of recollection should produce high pseudodiscrimination because *SP lures* would be mistaken for *targets*.

If list-length and list-strength manipulations act by specifically reducing recollection (assuming it is a recall-like process), we should observe an increase in pseudodiscrimination in *long* and *strong* lists compared to *short* lists. Increases in pseudodiscrimination for length and strength manipulations constitute *negative LLE* and *negative LSE*, respectively. Put another way, a negative LLE occurs when performance on *short* lists is *worse* than on *long* lists. Similarly, a negative LSE occurs when performance on non-strengthened items is *worse* on *short* lists than on *strong* lists. Thus, if the list manipulations affect recollection, negative LLEs and negative LSEs should be promptly observed.

Norman (2002, Exp. 2) found a negative LSE in SPU discrimination. The experiments in this thesis extend that methodology to list-length manipulations. The comparison between *SP lures* (*pseudotargets*) and *unrelated lures* can also be relevant to the discussion about the relative variances of lure distributions. Results showing  $z$ ROC slopes lower than 1 in SPU discrimination would suggest that the variance of the *SP lure* distribution is greater than that of *unrelated lures*. As a final note, the terms *SU / SSP / SPU comparison* and *SU / SSP / SPU discrimination* are used interchangeably throughout this thesis.

### 2.3.3. Power analysis

Power and effect sizes were calculated for most comparisons of interest in this thesis. Power is important in list-length and list-strength studies because of previous results showing null effects. It is crucial to assert whether or not the experimental designs used here have enough statistical power to detect previously reported interference effects. Effect sizes, on the other hand, are important not only because they are necessary to estimate power but also because they allow assessing the impact of different manipulations on the magnitude of interference effects. The power estimates and effect sizes reported in this thesis were calculated using G-Power 3.0 (Erdfelder, Faul, & Buchner, 1996; Faul, Erdfelder, Lang, & Buchner, 2007). Below we briefly describe those measures.

#### Power ( $1 - \beta$ )

Statistical power refers to the probability of detecting a significant difference between experimental conditions when there is in fact a difference. In general, hypothesis testing may lead to two types of error: *i*) Type I errors occur when a difference is claimed between conditions where in reality there is none; the probability ( $\alpha$ ) associated with this error is usually set before analysing the data; *ii*) Type II errors occur when no difference is claimed between conditions where in fact there is one; the probability ( $\beta$ ) associated with this error is dependent on  $\alpha$  and varies inversely with it, such that  $\beta$  increases when  $\alpha$  decreases. If  $\beta$  is the probability of missing a difference when there is one; its complement,  $1 - \beta$ , represents the probability of finding that difference. Thus,  $1 - \beta$  represents the power of the experiment.

Here we are interested in *post-hoc power* which refers to the assessment of the power of an experiment after it has been conducted. Post-hoc power enables estimating the chances of detecting a difference between conditions as large as the differences previously found in other experiments. Put another way, post-hoc power analyses enable estimating whether there is a reasonable chance of rejecting the null hypothesis of no effect given that there is a difference.

To estimate the power of an experimental comparison, three pieces of information are needed, namely, the  $\alpha$ -level of the comparison, the sample size and the size of the effect of interest in the population. The first two are readily available and the third can be estimated from sample data (see below). In our experiments,  $\alpha$  is set to .05. In addition, our  $\alpha$  value is two-tailed, meaning that the probability of falsely detecting an effect in either direction is actually halved. A natural consequence of this extra care is a decrease in power. On the other hand, power increases with increasing sample size and effect size. Thus, we strived to increase power in our experiments either by using many participants (e.g., Experiments 1, 2, 3, 7) or by increasing manipulation strength to boost effect sizes (e.g., Experiments 4, 6).

#### Effect size ( $d$ : independent samples $t$ -test)

A  $t$ -test is used to determine if two samples come from the same or different populations. The test assumes that the samples are independently drawn from Gaussian distributions. Population means ( $\mu_1, \mu_2$ ) and standard deviations ( $\sigma_1, \sigma_2$ ) are estimated from sample means ( $\bar{X}_1, \bar{X}_2$ ) and standard deviations ( $s_1, s_2$ ). The null hypothesis for inference is that there is no difference between the population means ( $H_0: \mu_1 - \mu_2 = 0$ ); the alternative hypothesis is that the means differ ( $H_1: \mu_1 - \mu_2 \neq 0$ ). The effect size  $d$  for an independent samples  $t$ -test is defined as:

$$d = \frac{(\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 + \sigma_2^2) / 2}} \quad (2.9)$$

If the standard deviations of the two populations are the same ( $\sigma_1 = \sigma_2$ ), then the denominator reduces to the common standard deviation  $\sigma$ . If standard deviations differ, then an average value (root mean square) is taken. Although  $t$ -tests assume equal standard deviations in both populations, they are nonetheless robust against violations of this assumption if sample sizes are similar ( $n_1 \approx n_2$ ).

In practice, effect sizes are estimated from sample means and standard errors of the mean. In these cases, standard deviations can still be estimated by using the relation ( $s = SEM \sqrt{2n_1n_2 / (n_1 + n_2)}$ ). When the sample sizes are equal, the relation reduces to  $s = SEM \sqrt{N}$ , where  $N$  is the common sample size. Thus, the

effect size for a given comparison in a published study can be estimated by taking the reported sample means, standard error of the means and sample sizes. The magnitude of effect sizes are conventionally classified as small ( $d = 0.2$ ), medium ( $d = 0.5$ ) and large ( $d = 0.8$ ) (Cohen, 1988).

#### Effect size ( $d_z$ : matched samples $t$ -test)

In studies where the same participant provides data for more than one condition, the sample data cannot be assumed to be *independently* drawn from a Gaussian distribution. That is because, for a given participant, performance in one condition is normally *correlated* to performance in another condition (i.e., a participant with a good memory tends to perform well in all conditions). As a result, a sample of  $N$  participants undergoing conditions  $x$  and  $y$  produces  $N$  pairs  $(x_i, y_i)$  of matched observations. To decide whether or not performance differs between the conditions, it is thus appropriate to frame the hypotheses in terms of the difference  $z_i = x_i - y_i$  between observations. The null hypothesis states that there is no difference between conditions ( $H_0: \mu_z = \mu_x - \mu_y = 0$ ); the alternative hypothesis states that there is a difference between conditions ( $H_1: \mu_z \neq 0$ ). The effect size  $d_z$  for a matched samples  $t$ -test is then defined as:

$$d_z = \frac{|\mu_z|}{\sigma_z} = \frac{|\mu_x - \mu_y|}{\sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y}} \quad (2.10)$$

where  $\mu_x$  and  $\mu_y$  are the means of populations  $x$  and  $y$ ,  $\sigma_x$  and  $\sigma_y$  are their corresponding standard deviations and  $\rho_{xy}$  is the correlation between the two conditions ( $\mu_z$  and  $\sigma_z$  are the mean and standard deviation of difference  $z$ ). As with  $d$  above,  $d_z$  can be estimated with the sample means, standard error of the means and sample sizes. Moreover, the correlation between the variables can be estimated with Pearson's correlation coefficient ( $r$ ). For studies where  $r$  is not reported, we assume  $r = .5$ . Most of the effect sizes reported in this thesis refer to matched pairs (within-participant) comparisons.

#### Effect size ( $g$ : ANOVA)

Analysis of Variance (ANOVA) is generally used to compare more than two experimental conditions at once. Most studies also provide data from pairwise comparisons between conditions, allowing the estimation of effect sizes with the



measures described above. However, some studies only provide data from the omnibus ANOVA across all conditions. In those cases, one can estimate effect sizes by comparing the highest and lowest mean values across conditions relative to the Mean Square Error (MSE), a measure of variance within each condition. This measure of effect size is called Hedges's  $g$  and is given by:

$$g = \frac{\mu_{\max} - \mu_{\min}}{\sqrt{MSE}} \quad (2.11)$$

where  $\mu_{\max}$  is the population mean with the highest value across conditions and  $\mu_{\min}$  is the mean with the lowest value. Hedge's  $g$  is related to Cohen's  $d$  by the relation  $d = g\sqrt{N / df}$ , where  $N$  is the total number of observations and  $df$  are the degrees of freedom of the error (i.e.,  $N - k - 1$ ; where  $k$  are the number of conditions in the experiment). For large sample sizes,  $N / df \approx 1$  and  $d \approx g$ .

#### 2.3.4. Regressions (zROC)

As described above, linear zROCs provide evidence of underlying Gaussian distributions. Some models, however, assume a threshold component, which predicts non-linear zROCs (e.g., Yonelinas, 2001). Thus, evaluating whether a straight line fits a zROC better than a non-straight line carries theoretical value (see 4.3.3 and 4.6.3 for a discussion of this issue in the context of our results).

First-order (linear) and second-order (quadratic) polynomials were fit to zROC data. Linear models ( $y = a_0 + a_1x$ ) had 2 parameters estimated ( $a_0, a_1$ ), whereas quadratic models ( $y = a_0 + a_1x + a_2x^2$ ) had 3 parameters estimated ( $a_0, a_1, a_2$ ).

Goodness of fit was measured with residual sum of squares  $RSS = \sum_{i=1}^n (o_i - p_i)^2$ ,

which represent the square of the difference between observed ( $o_i$ ) and predicted ( $p_i$ ) values at each of the  $n$  observed data points ( $n = 5$  points in the zROC).

Linear and quadratic models are nested because a linear model is simply a restricted version of a quadratic model in which  $a_2 = 0$ . Nested models can be compared against each other with a log-likelihood ratio test (Lamberts, 2005). If the difference in fits between the *general model* (quadratic) and the *restricted*

*model* (linear) is small, then the linear model is warranted as it captures the data pattern with fewer parameters. If, on the other hand, the difference in fits is large, then the linear model is not warranted as it lead to considerable loss of fit.

When goodness-of-fit is measured with *RSS*, the reliability of the difference in fits between general and restricted models is based on the  $\chi^2$  statistic given by

$$\chi^2 = -2 \ln \left[ \frac{RSS(\text{general})}{RSS(\text{restricted})} \right]^{\frac{n}{2}} \quad (2.12)$$

where  $n$  is the number of data points and the degrees of freedom of the test are given by the number of parameters eliminated from the general to the restricted model ( $df = 1$  here as only the quadratic term was removed from the general model). If the  $\chi^2$  statistic is greater than a critical value (for  $\alpha = .05$ ,  $\chi^2_{\text{crit}} = 3.84$ ), the restricted model is rejected as a description of the data in favour of the general version of the model.

## Chapter 3. Experiments 1-4

### 3.1. Introduction

In this chapter, we present four experiments designed to test some of the boundary conditions underlying the list-length and list-strength effects. The experiments described in this chapter follow closely the design adopted by Dennis and Humphreys' (2001, Exp. 2). List type (*short* vs. *long* vs. *strong* list) was manipulated within participants in Experiments 1 to 4.

One difference in the design adopted here is our use of a self-paced encoding task. Dennis and Humphreys (2001) used fixed (and long) encoding times (3 s). Long study times may allow participants to engage in rehearsal strategies (e.g., allocate some of the time during presentation of repeated items to rehearse non-repeated items). Rehearsal strategies may mask any existing list-length and list-strength effects. In an attempt to reduce the possible contribution of rehearsal, we used in Experiments 1 to 4 a self-paced encoding task in which participants were encouraged to move on to the next study item as quickly as possible.

In Experiment 1, encoding task was manipulated between participants. In Experiment 2, lure type was manipulated within participants. In Experiment 3, both encoding task (between participants) and lure type (within participants) were manipulated. Because retention interval in Experiment 3 was shorter than in Experiment 2, we also assessed the impact of retention interval on the magnitude of LLE and LSE by directly comparing the results of Experiments 2 and 3. Finally, in Experiment 4, both lure type (within participants) and retention interval (between participants) were manipulated. Because Experiment 4 also contained a more powerful manipulation (longer and stronger lists than in the previous experiments), we assessed the impact of manipulation strength on the magnitude of LLE and LSE by directly comparing the results of Experiments 3 and 4.

### 3.2. Experiment 1: Encoding task, long interval, 3x

In this experiment, participants performed either a size judgement encoding task (“does the item fit in the shoebox?”) or a pleasantness judgement task (“is the item pleasant?”). Norman (2002) found an LSE using the size task, whereas Dennis and Humphreys (2001) did not find an LSE using the pleasantness task. In the size task, participants have to encode items in an overlapping fashion, as they have to compare each item against the same referent (a shoebox present in the experimental room), whereas in the pleasantness task, they do not. It is thus possible that the type of processing an item receives at study mediates the amount of interference elicited by the other items on the list.

To test this possibility, we carried out an experiment in which encoding task was manipulated between participants and list type (*short*, *long* and *strong*) was manipulated within participants. Study time was self-paced and short (up to 3 s); retention interval was fixed and long (180 s); and strong items were presented three times (Dennis and Humphreys, 2001, Exp. 2). If the amount of encoding overlap at study is sufficient to elicit an LSE, then a list-strength effect should be observed in the size judgement condition but not in the pleasantness judgement condition. In other words, encoding task and list-strength should interact.

#### 3.2.1. Methods

##### Participants

Seventy-two University of Warwick students (25 males; age:  $M = 21.8$ ,  $SD = 3.7$ ) participated in the study (36 in the size judgement task and 36 in the pleasantness judgement task). The experiment lasted 45 min and participants were paid £5.

##### Materials

Stimuli were 360 imageable, concrete, familiar and medium-frequency nouns (Coltheart, 1981): mean imageability = 5.67 out of 7, range = 5.02-6.59; mean concreteness = 5.69 out of 7, range = 5.00-6.45; mean familiarity = 4.99, range = 4.00-6.16; mean Kučera-Francis frequency = 15.3 occurrences per million, range = 0-99; mean word length = 5.71, range = 3-10. The words were screened for

semantic similarity so that none of the items were strongly related to one another. This was achieved through pairwise matrix comparison using Latent Semantic Analysis (LSA; Landauer, Foltz, & Laham, 1998) with 300 feature dimensions applied on the GenCOL corpus, which is a sample of what a person would have read up to the first year at university.<sup>1</sup> Out of the initial pool of 1130 words, 360 were selected with cosine (a measure of semantic relatedness) less than 0.4.

Thirty words were used as fillers. The remaining 330 words were randomly assigned to 11 groups of 30 words, matched for word characteristics. Words were classified as *target* (if presented both at study and test), *interference* (if presented at study but not at test) or *lure* (if presented at test but not at study). Of the 11 groups, 3 consisted of targets, 5 consisted of interference words and 3 consisted of lures. A distinct word sample was produced for each participant so that, on average, the assignment of words to conditions was balanced.

## Design

Figure 3.1 illustrates the experimental design. Each participant attended one session. Each session consisted of four study/test blocks. The first block was practice. The three remaining blocks contained lists of three different types: *short list* (30 target items presented once and 30 interference items presented once), *long list* (30 target items presented once and 90 interference items presented once) and *strong list* (30 target items presented once and 30 interference items presented 3 times). List order was balanced across participants.

Participants were randomly assigned to one of two encoding conditions: *size judgement task* and *pleasantness judgement task*. In the *size* condition, subjects were given standard recognition memory task instructions and asked to decide

---

<sup>1</sup> LSA is used in this thesis as a rough measure of semantic relatedness and similarity mainly because previous memory studies have also used it (Norman, 2002; Howard & Kahana, 2002). For a given word pair, however, LSA has been shown not to be a good predictor of similarity ratings given by humans (Simmons & Estes, 2006). That is partly because LSA tends to assign large cosine values to antonyms (e.g., *black* vs. *white*), which are considered dissimilar by humans, and partly because the measure does not distinguish between taxonomic (feature-based) and thematic (relation-based) similarities, which are treated differently by humans. To minimise these problems, stimuli were further screened manually, allowing the elimination of words considered dissimilar by LSA but similar by humans. For example, *casket* and *coffin* were assigned a low cosine (.27) by LSA despite being synonyms; the word *coffin* was thus eliminated from the final stimulus set.

whether or not an item fits in a shoebox present in the experimental room (15 cm wide, 28 cm long, 10 cm deep). In the *pleasantness* condition, participants were asked to decide whether or not an item is pleasant.

List type	Size judgement task		
	Study	Distractor	Test
Short	( [AB] )	390 s	[tA, unr]
Long	( [AB] [CD] )	180 s	[tA, unr]
Strong	( [AB] [BB] )	180 s	[tA, unr]
	Pleasantness judgement task		
	Study	Distractor	Test
Short	( [AB] )	390 s	[tA, unr]
Long	( [AB] [CD] )	180 s	[tA, unr]
Strong	( [AB] [BB] )	180 s	[tA, unr]

**Figure 3.1. Design of Experiment 1.**

Judgement task was manipulated between participants; list type was manipulated within participants. In the size judgement condition, participants decided whether a typical instance of the study word would fit into a shoebox present in the experimental room. In the pleasantness judgement task, participants judged whether a typical instance of the study word was pleasant. Round brackets in the figure indicate that study time was self-paced: participants had up to 3 s to enter size or pleasantness judgements, meaning that study time could vary slightly across list types. A-D = matched groups of 30 words; [X,Y] = word groups X and Y were merged and the order of the resulting list was randomised; tA = targets were group A words; unr = unrelated lures.

In both conditions, a *retroactive design* was used: all target words were presented before any of the interference items were repeated. This prevents participants from telling targets from interference items, which potentially reduces differential rehearsal of targets during study.

Encoding condition (size vs. pleasantness judgement) was manipulated between participants (with 36 participants in each condition) and list type (*short*, *long* and *strong*) was manipulated within participants.

### Procedure

Stimuli were presented on a 43 cm CRT monitor. Each session consisted of four blocks: a practice block and three experimental blocks. Each block consisted of three phases: study, distractor and test.

*Study Phase.* Subjects were presented with 60 (*short*), 120 (*long*) and 120 (*strong*) items. Ten extra items were used as fillers (5 at the start and 5 at the end) of each study list to control for primacy and recency effects. Participants were warned that some items might appear several times. They were also informed that their memory would be tested. Participants were either instructed to decide whether or not a typical instance of the object denoted by the word would fit into a shoebox or whether the instance would be considered pleasant. Responses were entered on a 6-point rating scale ranging from *definitely yes* to *definitely no* by pressing the appropriate buttons on a gamepad. The task was self-paced, with an upper display time limit of 3,000 ms, after which the program automatically moved to the next item, and with an inter-stimulus interval of 500 ms.

*Distractor Phase.* A video game task was used to equate study-test lag across list types. The game, called “Eat the Squares”, required participants to use a gamepad in order to collect “green squares” randomly distributed on the screen and to avoid “deadly purple squares” also distributed on the screen. Each collected square added points to a counter at the bottom of the screen. Participants were instructed to accumulate as many points as possible. The game ended whenever a purple square was eaten. A new game then automatically started. This cycle was repeated for 390 s for *short* lists and 180 s for *long* and *strong* lists. These distractor times for *long* and *strong* lists were similar to the ones used by Dennis and Humphreys (2001): 4 min in their Experiment 2 compared to 3 min here.

*Test Phase.* The test list consisted of 60 words (30 old and 30 lures). Words appeared one at a time on the computer screen. Subjects were instructed to rate

their recognition confidence on a scale from 1 to 6 (*definitely old, probably old, guess old, guess new, probably new, definitely new*). They were encouraged to spread their ratings across the whole range of the scale. Response was self-paced.

### 3.2.2. Results

#### Hits and false alarms

A 2 (word type: *target, lure*)  $\times$  2 (encoding task: *size, pleasantness*)  $\times$  3 (list type: *short, long, strong*) mixed-design ANOVA on proportion of “old” responses revealed a main effect of list type,  $F(2,140) = 9.56$ ,  $MSE = 0.01$ ,  $p < .001$ , such that the proportion of “old” responses was lower in *strong* lists compared to *short* and *long* lists. There was also a main effect of encoding task,  $F(1,70) = 6.31$ ,  $MSE = 0.01$ , such that the proportion of “old” responses was lower in the *size* task compared to the *pleasantness* task. The interaction between word type and encoding task was marginally significant,  $F(1,140) = 3.43$ ,  $MSE = 0.01$ ,  $p = .07$ , indicating that hits decreased from the *pleasantness* to the *size* condition whereas false alarms remained unchanged.

Separate 2 (encoding task: *size, pleasantness*)  $\times$  3 (list type: *short, long, strong*) mixed-design ANOVAs were carried out on hits and false alarms. For hits, there was a main effect of encoding condition,  $F(1,70) = 16.68$ ,  $MSE = 0.01$ ,  $p < .001$ , such that hits in the *size* condition were lower than in the *pleasantness* condition. There was also a trend across list types, hinting that hits in the *strong* list were lower than hits in the *short* and *long* lists,  $F(2,140) = 2.07$ ,  $MSE = 0.01$ ,  $p = .13$ . There was no interaction between list type and encoding condition,  $F < 1$ ,  $p = .40$ . For false alarms, there was no effect of encoding condition,  $F < 1$ ,  $p = 0.77$ . There was, however, an effect of list type,  $F(2,140) = 7.75$ ,  $MSE = 0.01$ ,  $p = .001$ . Post-hoc LSD (Least Significant Difference) tests revealed that false alarms were lower for the *strong* list than for the *short* and *long* lists ( $ps < .001$ ); false alarms did not differ between *short* and *long* lists ( $p = .45$ ). There was no interaction between list type and encoding condition,  $F < 1$ ,  $p = .70$ . Hits and false alarms are presented in Table 3.1. Hits and false alarms broken down by encoding conditions are presented in Appendix 1 together with measures of sensitivity ( $d'$ ) and bias ( $c$ ).



There was no effect of length, as the interaction between word type (*target*, *lure*) and list type (*short*, *long*) was not significant,  $F < 1$ ,  $p = .44$ . There was also no effect of strength, as the interaction between word type (*target*, *lure*) and list type (*short*, *strong*) was not significant,  $F < 1$ ,  $p = .36$ . The fact that both hits and false alarms decreased with *strong* lists relative to *short* and *long* lists, suggests that participants adopted a more conservative response strategy with *strong* lists.

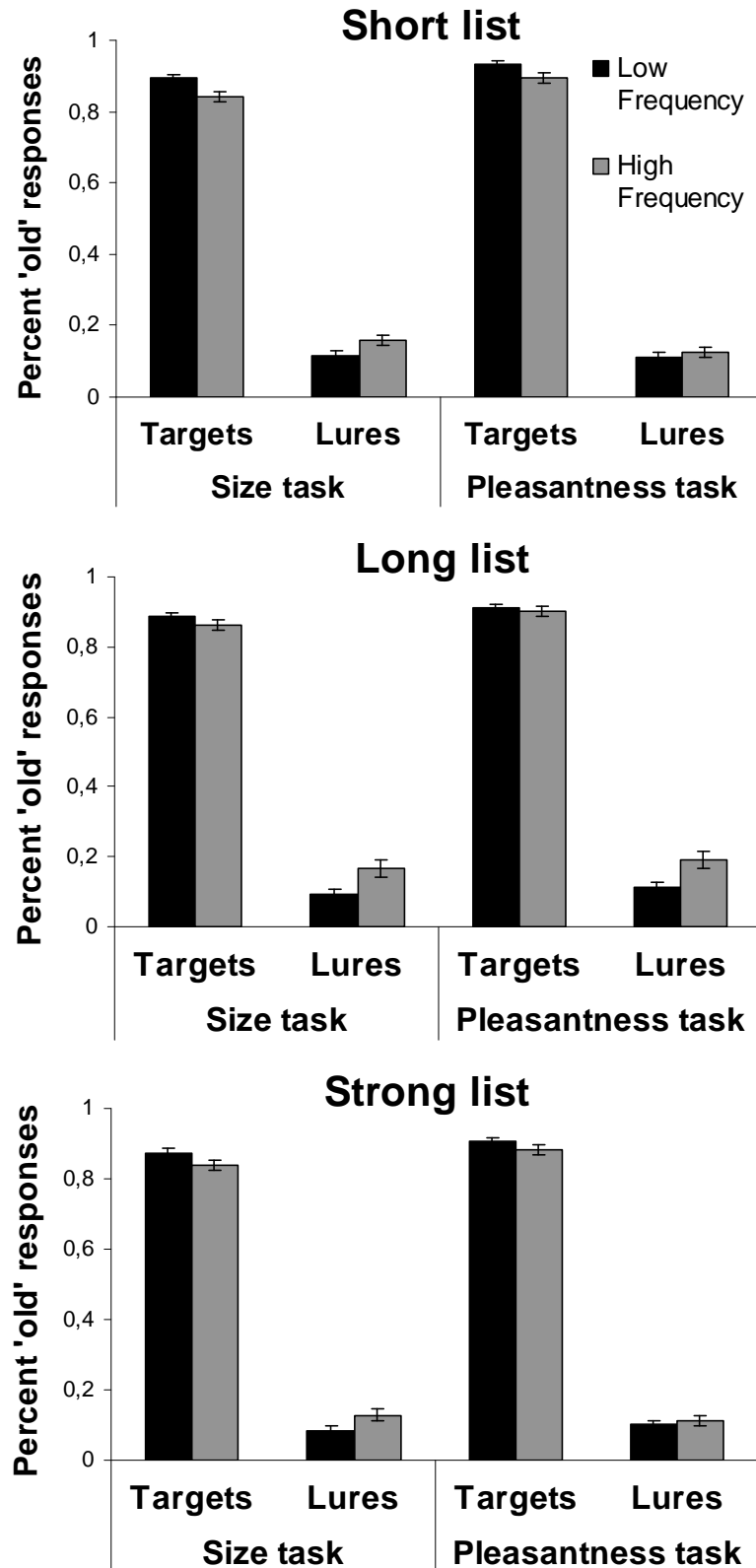
**Table 3.1. Hits and false alarms (Exp. 1).**

List type	HR Targets			FAR Unrelated lures		
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>
<b>Short</b>	.93	τ τ n *	.01	.11	τ τ n	.01
<b>Long</b>	.92	τ ⊥ n	.01	.12	τ ⊥ ***	.01
<b>Strong</b>	.91	⊥ ⊥	.01	.08	⊥ ⊥	.01

*Note.* HR = hits; FAR = false alarms. *n* non-significant; \*  $p \leq .05$ ; \*\*\*  $p \leq .001$ .  $N = 72$ . Data collapsed across encoding conditions (*size* and *pleasantness* judgement tasks).

To address the concern that the experiment may have lacked statistical power, we reanalysed the data in terms of word frequency. The null LLE and LSE in this experiment would be somewhat supported if observed in the context of a statistically significant effect (e.g., word-frequency mirror effect). Test words were split into low frequency (Kučera-Francis frequency  $< 4$ ; 117 words) and high frequency ( $\geq 20$ ; 85 words). A word-frequency mirror effect occurs if the interaction between word frequency (*low* vs. *high*) and word type (*target* vs. *lure*) is significant such that the proportion of “old” responses to *targets* decreases from *low*- to *high*-frequency words and the proportion of “old” responses to *lures* increases from *low*- to *high*-frequency words. Figure 3.2 summarises the results.

There was a word-frequency mirror effect in both *size* and *pleasantness* encoding conditions for *short* and *long* lists (all  $ps < .01$ ). For *strong* lists, the effect was smaller in the *size* condition ( $p = .02$ ) and non-significant in the *pleasantness* condition ( $p = .10$ ). The latter null result is consistent with studies showing that the mirror effect is reduced when participants are asked to carry out a pleasantness judgement task (Criss & Shiffrin, 2004b). These results give credence to the claim that Experiment 1 had enough power and that the lack of effects of list length and list strength may have been caused by factors other than insufficient power.



**Figure 3.2. Word-frequency effect across list types (Exp. 1).**

A *word-frequency mirror effect* occurs if there is a significant interaction between word frequency (*low* vs. *high*) and word type (*target* vs. *lure*) such that the proportion of “old” responses to *targets* decreases from *low*- to *high*-frequency words and the proportion of “old” responses to *lures* increases from *low*- to *high*-frequency words. The effect was found across encoding tasks and list types (all  $p$ s < .02), except for the *strong* list in the *pleasantness* task ( $p = .10$ ). Error bars = SEM.

### Sensitivity

A total of 216 Gaussian models (72 participants  $\times$  3 list types) were fitted to individual participants' confidence data; 4 were excluded due to poor fits. The results below refer to the estimates (sensitivity:  $A_z$ ; bias:  $c_a$ ) of the remaining 68 participants. Table 3.2 summarises the results collapsed across encoding tasks.

A 2 (encoding condition: *size*, *pleasantness*)  $\times$  3 (list type: *short*, *long*, *strong*) mixed-design ANOVA was carried out on the sensitivity measure ( $A_z$ ). There was no main effect of list type and no interaction between list type and encoding condition ( $F_s < 1$ ,  $p_s > .30$ ). There was, however, a main effect of encoding condition,  $F(1,66) = 6.03$ ,  $MSE = 0.01$ , showing that participants were worse at discriminating targets from lures in the *size* condition ( $M = .95$ ;  $SEM = .01$ ) than in the *pleasantness* condition ( $M = .96$ ;  $SEM = .01$ ). The difference in sensitivity across encoding tasks was driven mainly by the larger drop in hits in the *size* task.

**Table 3.2. Sensitivity ( $A_z$ ) and bias ( $c_a$ ) (Exp. 1).**

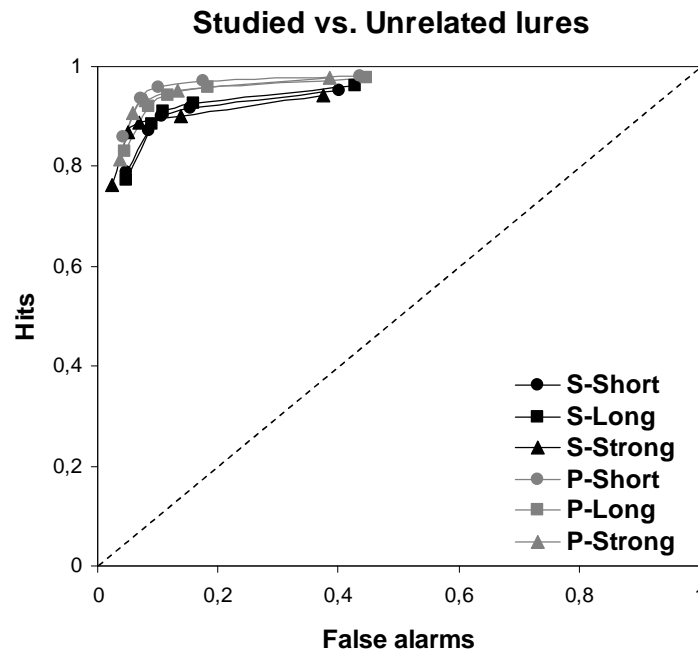
List type	$A_z$			$c_a$		
	$M$		$SEM$	$M$		$SEM$
<b>Short</b>	.96	$\tau$ $n$	$\tau$ $n$	.01	0.06	$\tau$ $n$ **
<b>Long</b>	.95	$\tau$ $n$	$\perp$ $n$	.01	0.07	$\tau$ $\perp$ **
<b>Strong</b>	.96	$\perp$ $n$	$\perp$ $n$	.01	0.17	$\perp$ $\perp$

*Note.*  $A_z$  = estimate of the area under the ROC;  $c_a$  = response bias (hits and false alarms obtained from the placement of the third criterion,  $X_3$ , on the familiarity space, separating *guess old* from *guess new* responses).  $n$  non-significant; \*\*  $p < .01$ . Data collapsed across encoding conditions (size and pleasantness judgement tasks).  $N = 68$ .

### Bias

A 2 (encoding condition: *size*, *pleasantness*)  $\times$  3 (list type: *short*, *long*, *strong*) mixed-design ANOVA on the bias measure ( $c_a$ ) revealed a main effect of list type,  $F(2,132) = 7.00$ ,  $MSE = 0.03$ ,  $p < .001$ ; LSD tests showed that participants were more conservative with *strong* lists than with *short* and *long* lists ( $p_s < .01$ ). There was also a marginal main effect of encoding task,  $F(1,66) = 13.28$ ,  $MSE = 0.15$ ,  $p = .07$ , suggesting that participants in the *size* condition were more conservative than participants in the *pleasantness* condition. There was no interaction between list type and encoding task ( $p = .28$ ). Figure 3.3 shows the ROC curves pooled across participants for each list type and encoding task. The curves from the *size*

condition are shifted downwards and leftwards relative to the curves from the *pleasantness* condition, indicating lower sensitivity and higher bias.



**Figure 3.3. ROC curves for Experiment 1.**

Lists learned in the *size* task were worse recognised than those learned in the *pleasantness* task, despite near-ceiling performance. Curves from the *size* condition were shifted down and to the left, indicating lower sensitivity and higher bias. S = size task ( $N = 34$ ); P = pleasantness task ( $N = 34$ ).

### 3.2.3. Discussion

The results of Experiment 1 show that discrimination in the *size* judgement task was impaired relative to discrimination in the *pleasantness* judgement task and that this difference was driven by a reduction in hits in the *size* condition rather than by an increase in false alarms. Not only sensitivity was lower in the *size* condition but also bias was more conservative and retrieval times were longer in that condition (see Appendix 2). That is, participants were less willing to endorse a particular test word as “old” and thus took longer to make a recognition decision when items were studied with the *size* judgement task. The differences in bias and response times between encoding tasks were probably not caused by differences in task difficulty as the encoding times were the same across tasks. In other words, it is unlikely that participants discriminated less well between targets and lures, were more reluctant to name an item “old” and took longer to respond simply because they did not encode the study words as well in the *size* condition. However, the

interpretation of response times here should not be taken too seriously as emphasis was given to accuracy over speed in the instructions given to participants.

Despite the overall effect of encoding task on sensitivity, no difference was observed across list types (i.e., no LLE and no LSE). The null result suggests that the different encoding tasks used by Dennis and Humphreys (2001) and Norman (2002) was not the critical factor underlying their discrepant results. Moreover, the results indicate that the type of processing an item receives at study may not, by itself, appreciably increase the amount of interference elicited by the other items on the list. If it were so, an interaction between list type and encoding task should have been observed, such that LLE and LSE should have been found in the *size* judgement condition but not in the *pleasantness* condition. Thus, at face value, the null results obtained here provide support for the BCDMEM model, which poses that interference effects in recognition are caused by noise from other contexts in which a word was studied rather than by noise from other items in a study list.

It could be argued that the length and strength manipulations used here were not strong enough to elicit detectable effects. For example, how do we know the strong items were indeed strengthened if they were never tested? The evidence that the strength manipulation was effective, although indirect, comes from the large changes in bias specific to the *strong* lists. Hirshman (1995, p. 306) pointed out that criterion shifts in mixed-strength lists only occur when there are large and statistically significant differences in sensitivity between weak and strong items (for example, when the difference in  $d'$  is greater than 0.5). Thus, the large increase in bias found here suggests that the strength manipulation was effective. As to the list-length manipulation, previous research has found reliable effects with long-to-short length ratios as low as the one used in the present experiment (e.g., Criss & Shiffrin, 2004c, Exp. 1; Zaki & Nosofsky, 2001, Exp. 2), suggesting that the 2:1 ratio used here might be appropriate to elicit a list-length effect.

There are, however, other reasons to believe that the lack of LLE and LSE in this study should not obtain generally. Cary and Reder (2003) found an LLE using Dennis and Humphreys' (2001) design, which is similar to the design used here, and Norman (2002, Exp. 1) found an LSE using only unrelated lures. Those

studies, however, used manipulations more powerful than the manipulations used here (4:1 vs. 2:1 long-to-short ratio; 6 vs. 3 word repetitions). We address the impact of those procedural differences in Experiment 4. For the next two experiments, we kept the length ratios and number of repetitions as they are and instead varied the type of lures present at test (Experiment 2) and both the type of lures at test and the encoding task (Experiment 3).

### 3.3. Experiment 2: Lure type, long interval, 3x

The failure to find an LLE and an LSE in Experiment 1 may have been caused by the small contribution of recollection at test. Dual-process models, such as SAC and CLS, predict that length and strength effects occur when recollection is selectively impaired (though familiarity may also be affected). If participants were relying mostly on familiarity to base their recognition decisions, and if familiarity is relatively less harmed by interference, then negligible effects should ensue.

In order to increase the relative contribution of recollection at test, we carried out an experiment similar to Experiment 1 with two exceptions. First, all participants took part in a *size* judgement encoding task. Second, both related lures (*SP lures*, study: *banana*, test: *bananas*) and unrelated lures (test: *car*) were presented at test.

The *size* judgement task was chosen (as opposed to the *pleasantness* task) because it allows performance to drop below ceiling, thereby facilitating the detection of differences, if any, between list types. Moreover, the task has been successfully used in previous studies where LSEs have been observed (Norman, 1999, 2002).

Norman (2002, Exp. 2) hypothesised that an LSE should be observed when recollection plays the main role at test but not when familiarity is the leading process. He implemented this idea by comparing discrimination between targets and SP lures (taken to be an index of recollection) and discrimination between targets and unrelated lures (taken to be an index of familiarity). Consistent with the hypothesis, he found an LSE in SSP comparisons (Studied items vs. Switched-Plurality lures) but not in SU comparisons (Studied items vs. Unrelated lures).

Participants were less able to recall a lure as “new” when their corresponding targets were studied in *mixed* lists (i.e., weak items in *strong* lists) than when their corresponding targets were studied in weak lists. By contrast, participants’ ability to discriminate between targets and unrelated lures was unaffected by list type.

The use of SP lures may also shed light on the dependency of LLE on recollection. Studies using the process-dissociation procedure have shown that the recollection estimate (but not the familiarity estimate) is affected by list-length manipulations (Yonelinas, 1994; Yonelinas & Jacoby, 1994). Moreover, both SAC and CLS models, within certain parameter ranges, predict that increases in list length should selectively decrease recollection.

It is hypothesised that length and strength manipulations should selectively reduce discrimination between targets and SP lures but not between targets and unrelated lures. To increase the engagement in recall-to-reject at test, participants were told that recalling a word in its singular (or plural) form meant that the plural (or singular) form of that word had not been studied and, therefore, that a *definitely new* response should be entered (see Rotello et al., 2000, Exps. 1 and 2 ).

### 3.3.1. Methods

#### Participants

One-hundred and twenty-six University of Warwick student (67 males; age:  $M = 21.7$ ,  $SD = 3.8$ ) participated in the study. Participants were tested individually. Each session took about 45 minutes and participants were paid £5.

#### Materials

Stimuli were 360 imageable, concrete, familiar and medium-frequency nouns from the MRC Psycholinguistic Database: mean imageability = 5.69 out of 7, range = 5.02-6.59; mean concreteness = 5.72 out of 7, range = 5.00-6.48; mean familiarity = 5.04, range = 4.00-6.16; mean Kučera-Francis frequency = 15.88 occurrences per million, range = 0-99; mean word length = 5.62, range = 3-10.<sup>2</sup>

---

<sup>2</sup> Word properties were slightly different between Experiments 1 and 2. In the latter, words needed to be such that their plural forms were created by adding an *s* to their singular forms. To satisfy this constraint in Experiments 2, some words from Experiment 1 had to be replaced.

The words were screened for semantic relatedness as in Experiment 1 (see 3.2.1). Thirty words were used as fillers. The remaining 330 words were randomly assigned to 11 groups of 30 words, matched for word characteristics. Words were classified as target, interference or lure. The lures were further classified as *SP lures* (switched-plurality; e.g., study *banana*, test *bananas*) or *unrelated lures* (e.g., study *banana*, test *car*). Of the 11 word groups, 3 consisted of targets, 5 consisted of interference words and 3 consisted of unrelated lures. SP lures were constructed by switching the plurality of half the targets (from singular to plural or vice-versa). Plural forms were generated by adding an *s* to their singular forms. Distinct samples were produced for each participant.

### Design and Procedure

Figure 3.4 illustrates the experimental design. Design and Procedure were identical to Experiment 1 with three exceptions. First, only the size judgement task was used during encoding. Second, switched-plurality lures were used at test in addition to unrelated lures. Third, participants were told to try and recall studied words whenever possible (e.g., when presented with highly familiar SP lures). The test list consisted of 60 words (15 old items, 15 SP lures and 30 unrelated lures) and response was self-paced.

List type	Study	Distractor	Test
Short	( [AB] )	390 s	[tA, spA, unr]
Long	( [AB] [CD] )	180 s	[tA, spA, unr]
Strong	( [AB] [BB] )	180 s	[tA, spA, unr]

**Figure 3.4. Design of Experiment 2.**

Participants were told to decide whether a typical instance of the study word would fit into a shoebox. A-D = groups of 30 words; [X,Y] = word groups X and Y were merged and the order of the resulting list was randomised; tA = targets were half of the words from group A; spA = switched-plurality lures were the other half of the words from group A; unr = unrelated lures.



### 3.3.2. Results

#### Hits and false alarms

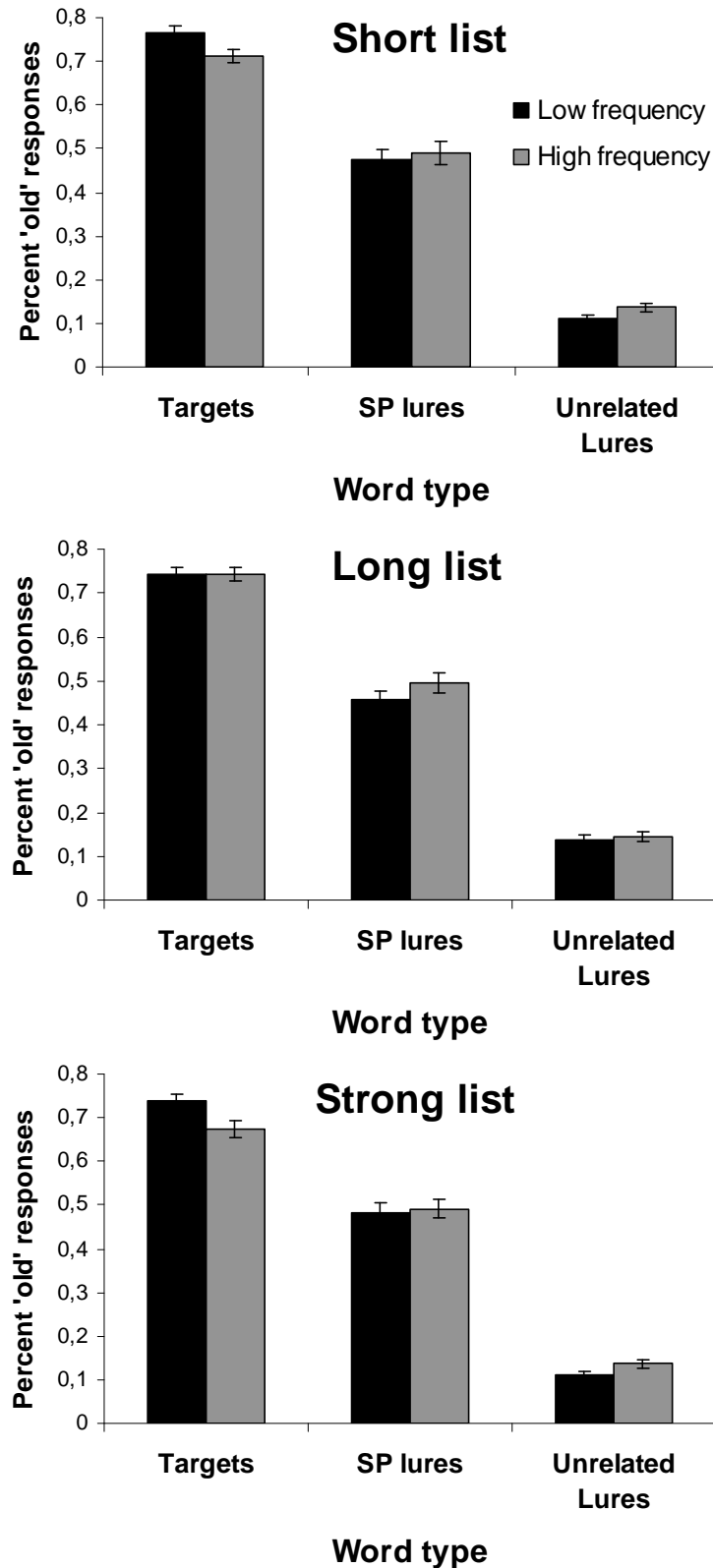
Separate one-way repeated-measures ANOVAs were carried out on the proportion of “old” responses for each word type (*target*, *SP lure* and *unrelated lure*) with list type (*short*, *long* and *strong*) as the independent variable. For *targets*, there was an effect of list type,  $F(2,250) = 3.04$ ,  $MSE = 0.01$ ,  $p = .05$ , such that hit rates were lower for *strong* lists compared to *short* and *long* lists ( $p \leq .05$ ) but did not differ between *short* and *long* lists ( $p = .74$ ). For *SP lures*, there was no difference across list types,  $F < 1$ ,  $p = .94$ . For *unrelated lures*, there was an effect of list type,  $F(2,250) = 4.54$ ,  $MSE = 0.01$ ,  $p = .01$ , such that false alarms to unrelated lures were higher for *long* lists than for both *short* ( $p = .04$ ) and *strong* lists ( $p < .01$ ). When word type and list type were entered into the same repeated-measures ANOVA, there also was no interaction between the two variables,  $F < 1$ ,  $p = .50$ . Hits and false alarms are presented in Table 3.3. Sensitivity ( $d'$ ) and bias ( $c$ ), with lure types analysed either separately or together, are reported in Appendix 1.

**Table 3.3. Hits and false alarms (Exp. 2).**

List type	HR Targets			FAR SP lures			FAR Unrelated		
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>
<b>Short</b>	.80	$\tau$ $n$	.01	.45	$\tau$ $n$	.02	.10	$\tau$ $*$	.01
<b>Long</b>	.81	$\tau$ $*$	.01	.46	$\tau$ $n$	.02	.12	$\tau$ $*$	.01
<b>Strong</b>	.78	$\tau$ $*$	.01	.45	$\tau$ $*$	.02	.10	$\tau$ $**$	.01

*Note.* SP = switched plurality;  $n$  non-significant;  $*$   $p \leq .05$ ;  $**$   $p < .01$ .  $N = 126$ .

There was no effect of length, as the interactions between word type (*target* vs. *SP lure*; *target* vs. *unrelated lure*) and list type (*short* vs. *long*) were not significant,  $F_s < 1$ ,  $p_s > .40$ . There was also no effect of strength, as the interactions between word type (*target* vs. *SP lure*; *target* vs. *unrelated lure*) and list type (*short* vs. *strong*) were not significant,  $F_s < 2.1$ ,  $p_s > .15$ . Hits and false alarms, however, changed in ways consistent with harmful effects of list length and strength. False alarms increased with list length (while hits remained unchanged) and hits decreased with list strength (while false alarms remained unchanged).



**Figure 3.5. Word-frequency effect across list types (Exp. 2).**

The interaction between word frequency (*low* vs. *high*) and word type (*target* vs. *lure*), which indexes the word-frequency mirror effect, was significant for both *SP* and *unrelated* lures and for both *short* and *strong* lists (all  $ps < .04$ ), but not for *long* lists ( $ps > .20$ ). Error bars = SEM.

As in Experiment 1, we reanalysed the data in terms of word frequency to provide some evidence against the criticism that the experiment lacked power. Low-frequency (Kučera-Francis frequency  $< 4$ ; 98 words) and high-frequency (Kučera-Francis frequency  $\geq 20$ ; 77 words) words were analysed separately for each lure type (*SP* vs. *unrelated*) and each list type (*short*, *long* and *strong*) with word frequency (*low* vs. *high*) and word type (*target* vs. *lure*) as the independent variables. A word-frequency mirror effect was found in all cases (all  $ps < .04$ ), except for *long* lists ( $ps > .20$ ). Again as in Experiment 1, the mirror effects were obtained despite unfavourable conditions: not only has it been previously shown that the use of encoding tasks attenuates the hit-rate portion of the effect (Criss & Shiffrin, 2004b) but it has also been shown that the false-alarm portion of the effect is reduced when SP lures are used (Arndt & Reder, 2002). Thus, at least for *short* and *strong* lists, the experimental design was powerful enough to produce reliable word-frequency mirror effects. Figure 3.5 summarises the results.

### Sensitivity

Sensitivity  $A_z$  was estimated by fitting Gaussian models to participants' confidence data. Of the 1134 models fitted (126 participants  $\times$  3 list types  $\times$  3 comparison types<sup>3</sup>: SU, SSP and SPU), 13 were excluded due to poor fits ( $\chi^2 p$ -value  $< .05$ ); this screening addresses the concern that  $A_z$  may misrepresent sensitivity when the model is a poor fit. The results refer to the sensitivity and bias estimates of the remaining 114 participants whose data were reasonably fitted by the model across all three discrimination types<sup>4</sup>. Table 3.4 summarises the results.

**Table 3.4. Sensitivity ( $A_z$ ) across discrimination types (Exp. 2).**

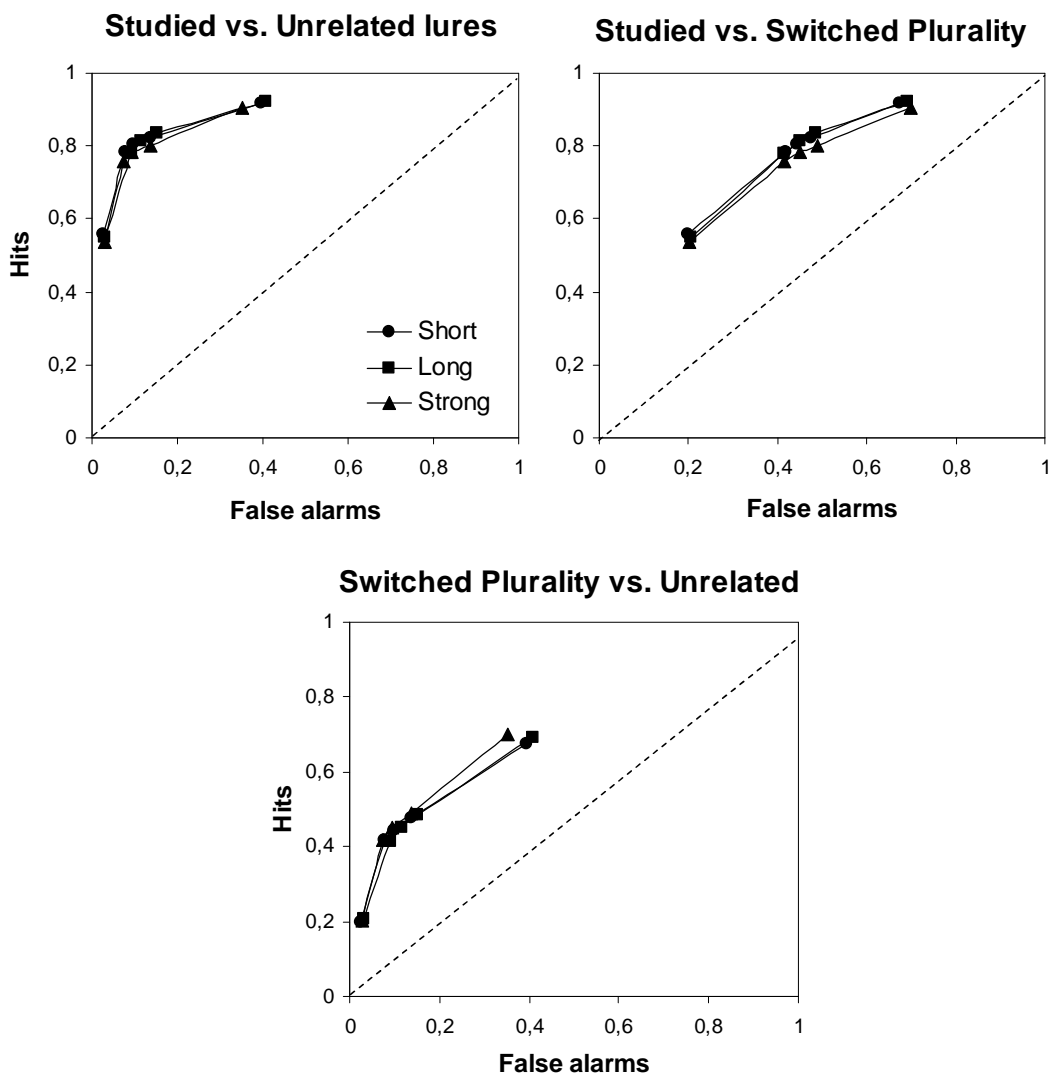
List type	SU			SSP			SPU		
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>
<b>Short</b>	.90	$\tau$ $n$ $\perp$	.01	.74	$\tau$ $n$ $\perp$	.01	.72	$\tau$ $n$ $\perp$	.02
<b>Long</b>	.90	$\tau$ $\perp$	.01	.73	$\tau$ $\perp$	.01	.72	$\tau$ $\perp$	.02
<b>Strong</b>	.90	$n$ $\perp$	.01	.73	$n$ $\perp$	.01	.77	** $\perp$	.01

*Note.*  $A_z$  = area under the ROC; SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality lures; SPU = switched-plurality vs. unrelated. *n* non-significant; \*\*  $p < .01$ .

<sup>3</sup> See 2.3.2 for an explanation of the comparison types SU, SSP and SPU.

<sup>4</sup> Note that the same participant may produce more than one poor fit (3 lists  $\times$  3 comparison types). Thus the number of rejected models may be higher than the number of rejected participants.

Separate one-way repeated-measures ANOVAs were performed on the sensitivity measure ( $A_z$ ) for each discrimination type (SU, SSP and SPU) with list type (*short*, *long* and *strong*) as the independent variable. There was no difference across list types in SU and SSP comparisons ( $F_s < 1$ ,  $p_s > .66$ ). In the SPU comparison, however, there was a significant difference in *pseudodiscrimination* across list types,  $F(2,226) = 6.32$ ,  $MSE = 0.01$ ,  $p < .01$ . Post-hoc LSD tests showed that participants were less likely to correctly reject SP lures in *strong* lists than in both *short* and *long* lists ( $p < .01$ ). The rise in pseudodiscrimination with *strong* lists, suggests a small LSE. Overall, however, the results show no LLE and no LSE.



**Figure 3.6. ROC curves for Exp. 2.**

In the “Studied vs. Switched Plurality” comparison, the curve for the *strong* list lies below the curves for both *short* and *long* lists (non-significant difference), whereas in the “Switched Plurality vs. Unrelated” comparison, the *strong* curve lies above the other curves ( $p < .01$ ;  $N = 114$ ).

Figure 3.6 shows ROC curves pooled across participants for each list type and discrimination type. In the SSP comparison, the curve for the *strong* list lies below the curves for both *short* and *long* lists, whereas in the SPU comparison, the *strong* curve lies above the other curves.

The null effects observed here are unlikely to be a result of low power. Using the  $A_z$  values reported by Norman (2002, Exp. 2, SSP comparison) to compute the size of his LSE ( $d_z = 0.44$ ), we found that the estimated power to detect an LSE of that size in our Experiment 2 ( $N = 114$ ) is .99. Also, the power to detect an LSE half the size of the effect found by Norman (2002) is .75, which is close to the traditional .80 power threshold adopted in most studies (Cohen, 1988).

### Bias

Separate one-way ANOVAs were carried out on the bias measure ( $c_a$ ) for each discrimination type (SU, SSP and SPU) with list type (*short*, *long* and *strong*) as the independent variable. In the SU comparison, there was a difference in bias across list types,  $F(2,226) = 6.20$ ,  $MSE = 0.05$ ,  $p < .01$ : participants were more conservative in *strong* lists than in both *short* ( $p = .02$ ) and *long* ( $p = .001$ ) lists and equally conservative in *short* and *long* lists ( $p = .24$ ). In the SSP comparison, there was no difference across lists,  $F < 1$ ,  $p = .38$ . In the SPU comparison, there was a marginal difference across lists,  $F(2,226) = 2.60$ ,  $MSE = 0.06$ ,  $p = .08$ ; participants were more conservative in *strong* lists. Table 3.5 shows these results.

**Table 3.5. Bias ( $c_a$ ) across discrimination types (Exp. 2).**

List type	SU			SSP			SPU		
	$M$		$SEM$	$M$		$SEM$	$M$		$SEM$
<b>Short</b>	0.21	$\tau$ $n$ $\perp$	0.02	-.33	$\tau$ $n$ $\perp$	0.03	0.65	$\tau$ $n$ $\perp$	0.03
<b>Long</b>	0.18	$\tau$ $\perp$	0.03	-.36	$\tau$ $\perp$	0.04	0.63	$\tau$ $\perp$	0.04
<b>Strong</b>	0.28	** $\perp$	0.03	-.31	$n$ $\perp$	0.04	0.71	* $\perp$	0.04

*Note.*  $c_a$  = response bias; SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality lures; SPU = switched-plurality vs. unrelated.  $n$  non-significant;  $\dagger p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ .

### 3.3.3. Discussion

The results of Experiment 2 show that discrimination does not change across list types for unrelated lures (SU comparison) but decreases slightly, though non-significantly, for SP lures in *long* and *strong* lists (SSP comparison). By contrast, *pseudodiscrimination* (SPU comparison) was significantly higher for *strong* lists. As in Experiment 1, participants were more conservative in their “old” responses to *strong* lists, suggesting that the strength manipulation was effective (see 3.2.3). The combination of higher bias and higher pseudodiscrimination in *strong* lists provides some evidence of an LSE: despite responding “old” less often in *strong* lists, participants still responded “old” more often to SP lures in *strong* lists than in both *short* and *long* lists. The overall data pattern is thus consistent with the results reported by Norman (2002, Exp. 2), hinting that list-strength manipulations are more harmful to recollection (SSP) than to familiarity (SU).

The results, however, are also consistent with the null results reported by Dennis and Humphreys (2001), since no significant difference in discrimination across lists was found here in the SSP comparison. Power analysis suggested that the null results were probably not caused by low statistical power. Thus, at the very least, the results of Experiment 2 indicate that forcing participants to use recollection at test is not sufficient to elicit fully fledged list-length and list-strength effects.

## 3.4. Experiment 3: Lure type, short interval, 3x, enc. task

The manipulations of list length and list strength in Experiments 1 and 2 have failed to produce significant interference effects. Although it is possible that longer lists and more word repetitions are needed to reach detectable levels of interference (e.g., Norman, 2002; Cary and Reder, 2003), another possibility is that the long retention interval between the end of the study list and the beginning of the test list (180 s) may have contributed to the lack of an effect.

Both item-noise and context-noise models predict that retention interval should modulate interference effects. For item-noise models, interference between list items should be reduced because memory strength decreases with retention

interval (Gehring, Toggia, & Kimble, 1976; Strong, 1913); if traces become weaker, they are less likely to interfere with one another. For context-noise models, interference should be reduced with longer retention intervals because this extra time presumably facilitates the reinstatement of the original study context. This contrasts to short retention intervals, where contextual inertia may prevent the reinstatement of an appropriate context, resulting in poorer performance.

To test the modulatory role of retention interval on interference, we repeated Experiment 2 with two differences. First, retention interval for *long* and *strong* lists was reduced from 180 s to 0 s. Second, encoding task (*size* vs. *pleasantness*) was manipulated. The encoding task manipulation was included in order to test whether the previous lack of interaction between encoding task and list type in Experiment 1 was simply a consequence of its longer retention interval.


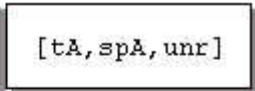

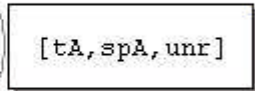

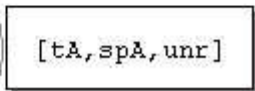
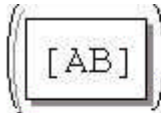
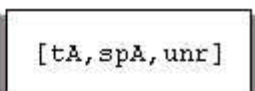
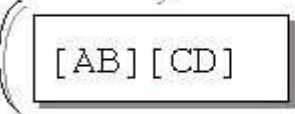
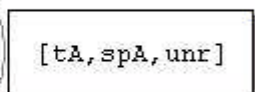

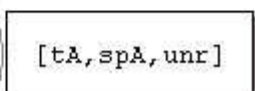
### 3.4.1. Methods

#### Participants

One-hundred and thirty-two University of Warwick students (53 males; age:  $M = 21.5$ ,  $SD = 4.4$ ) participated in the study: 66 took part in the size judgement encoding task and 66 in the pleasantness judgement task. Participants were tested individually. Each session took 40 minutes and participants were paid £5.

#### Materials, Design and Procedure

Figure 3.7 illustrates the experimental design. Materials were identical to Experiment 2. Design and Procedure were identical to Experiment 2, except that participants were randomly assigned to one of two encoding conditions (*size* judgement or *pleasantness* judgment task) and that distractor time was reduced across list types (210 s for *short* lists and 0 s for *long* and *strong* lists). With no retention interval for *long* and *strong* lists (0 s), the detrimental effect of interference items, which come late in the study list, should be maximal, as those items would still be freshly represented in memory. Encoding condition (*size* vs. *pleasantness*) was manipulated between participants (with 66 participants in each condition) and list type (*short*, *long*, *strong*) was manipulated within participants.

List type	Size judgement task		
	Study	Distractor	Test
Short		210 S	
Long			
Strong			
	Pleasantness judgement task		
	Study	Distractor	Test
Short		210 S	
Long			
Strong			

**Figure 3.7. Design of Experiment 3.**

Judgement task was manipulated between participants; list type was manipulated within participants. A-D = groups of 30 words; [X,Y] = word groups X and Y were merged and the order of the resulting list was randomised; tA = targets were half of the words from group A; spA = switched-plurality lures were the other half of the words from group A; unr = unrelated lures.

### 3.4.2. Results

#### Hits and false alarms

A 3 (word type: *target*, *SP lure*, *unrelated lure*)  $\times$  2 (encoding task: *size*, *pleasantness*)  $\times$  3 (list type: *short*, *long*, *strong*) mixed-design ANOVA on proportion of “old” responses revealed a marginal main effect of list type,  $F(2,260) = 2.48$ ,  $MSE = 0.01$ ,  $p = .08$ , such that the proportion of “old” responses was lower in *strong* lists than in *short* and *long* lists. There was also a marginal main effect of encoding task,  $F(1,130) = 3.11$ ,  $MSE = 0.08$ ,  $p = .08$ , such that the proportion of “old” responses was lower in the *size* task than in the *pleasantness* task. The interaction between word type and encoding task was not significant,



$F(2,260) = 2.14$ ,  $MSE = 0.05$ ,  $p = .12$ ; the means, however, showed a trend consistent with the results of Experiment 1, whereby hits were lower in the *size* condition than in the *pleasantness* condition, whereas false alarms to unrelated lures were almost identical across encoding conditions. The interaction between word type and list type was significant,  $F(4,520) = 2.97$ ,  $MSE = 0.01$ , such that hits and false alarms to unrelated lures were lower to *strong* lists than to *short* and *long* lists but false alarms to SP lures did not differ across list types.

Separate 2 (encoding condition: *size*, *pleasantness*)  $\times$  3 (list type: *short*, *long*, *strong*) mixed-design ANOVAs were performed on proportion of “old” responses for each word type (*targets*, *SP lures* and *unrelated lures*). For hits, there was a main effect of list type,  $F(2,260) = 3.74$ ,  $MSE = 0.01$ , showing that hits were lower in *strong* lists than in both *short* and *long* lists ( $p = .02$ ) but did not differ between *short* and *long* lists ( $p = .91$ ). There was no main effect of encoding task. There was, however, an interaction between encoding task and list type,  $F(2,260) = 5.06$ ,  $MSE = 0.01$ ,  $p < .01$ , such that hits were lower in the *size* condition than in the *pleasantness* condition for *short* and *strong* lists but not for *long* lists. For false alarms to SP lures, there was no main effect of list type and no interaction,  $F_s < 1$ ,  $p_s > .53$ , but there was a marginal main effect of encoding task,  $F(1,130) = 3.18$ ,  $MSE = 0.14$ ,  $p = .08$ , suggesting that SP false alarms were overall lower in the *size* condition. Finally, for unrelated lures there was a main effect of list type,  $F(2,260) = 9.89$ ,  $MSE = 0.01$ ,  $p < .001$ : false alarms were lower in *strong* lists than in *short* and *long* lists ( $p_s < .001$ ) and did not differ between *short* and *long* lists ( $p = .63$ ). There was no effect of encoding condition and no interaction,  $F_s < 1$ ,  $p_s > .47$ .

**Table 3.6. Hits and false alarms across encoding tasks (Exp. 3).**

List type	HR Targets			FAR SP lures			FAR Unrelated		
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>
<b>Short</b>	.80	$\tau$ $n$	.01	.46	$\tau$ $n$	.02	.11	$\tau$ $n$	.01
<b>Long</b>	.81	$\tau$ $*$	.01	.47	$\tau$ $n$	.02	.10	$\tau$ $***$	.01
<b>Strong</b>	.77	$\perp$ $*$	.01	.48	$n$ $\perp$	.02	.08	$\perp$ $\perp$	.01

*Note.* SP = switched plurality; *n* non-significant; \*  $p \leq .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .  $N = 132$ . Data collapsed across encoding conditions (size and pleasantness judgement).

Hits and false alarms, collapsed across encoding tasks, are presented in Table 3.6. Mean sensitivity ( $d'$ ) and bias ( $c$ ) values for each encoding condition and with lure types analysed either separately or together are reported in Appendix 1.

There was no effect of length, as the interactions between word type (*target* vs. *SP lure*; *target* vs. *unrelated lure*) and list type (*short* vs. *long*) were not significant,  $F_s < 1$ ,  $p_s > .70$ . There was also no effect of strength when comparing hits with unrelated lures, as the interaction between word type (*target* vs. *unrelated lure*) and list type (*short* vs. *strong*) was not significant,  $F_s < 1$ ,  $p = .95$ . There was, however, an effect of strength when comparing hits with SP lures, as the interaction between word type (*target* vs. *SP lure*) and list type (*short* vs. *strong*) was significant,  $F(1,131) = 4.81$ ,  $MSE = 0.01$ ; the latter interaction showed that hits decreased and false alarms increased from *short* to *strong* lists.

We also reanalysed the data from *long* lists in terms of word frequency to assess whether the experimental design was powerful enough to elicit this robust effect. The  $2$  (word frequency: *low*, *high*)  $\times 2$  [word type: *target*, *lure* (*SP* + *unrelated*)] repeated-measures ANOVA revealed a large word-frequency mirror effect,  $F(1,130) = 11.21$ ,  $MSE = 0.02$ ,  $p = .001$ , showing that hits decreased (.76 vs. .71) and false alarms increased (.25 vs. .28) from *low* to *high* frequency words. The mirror effect was significant for both *size* and *pleasantness* judgement tasks.

All in all, the raw data indicate that participants responded “old” less frequently in the *size* condition than in the *pleasantness* condition. Moreover, hits and false alarms did not interact in a way consistent with an effect of list length. Finally, hits and false alarms behaved in a way consistent with a harmful effect of list strength on memory but only when *targets* were compared to *SP lures*.

### Sensitivity

Of the 1188 Gaussian models fitted to participants’ data (132 participants  $\times$  3 list types  $\times$  3 comparison types: SU, SSP and SPU), 17 were excluded due to poor fits ( $\chi^2$   $p$ -value  $< .05$ ). The data below refers to the remaining 119 participants.

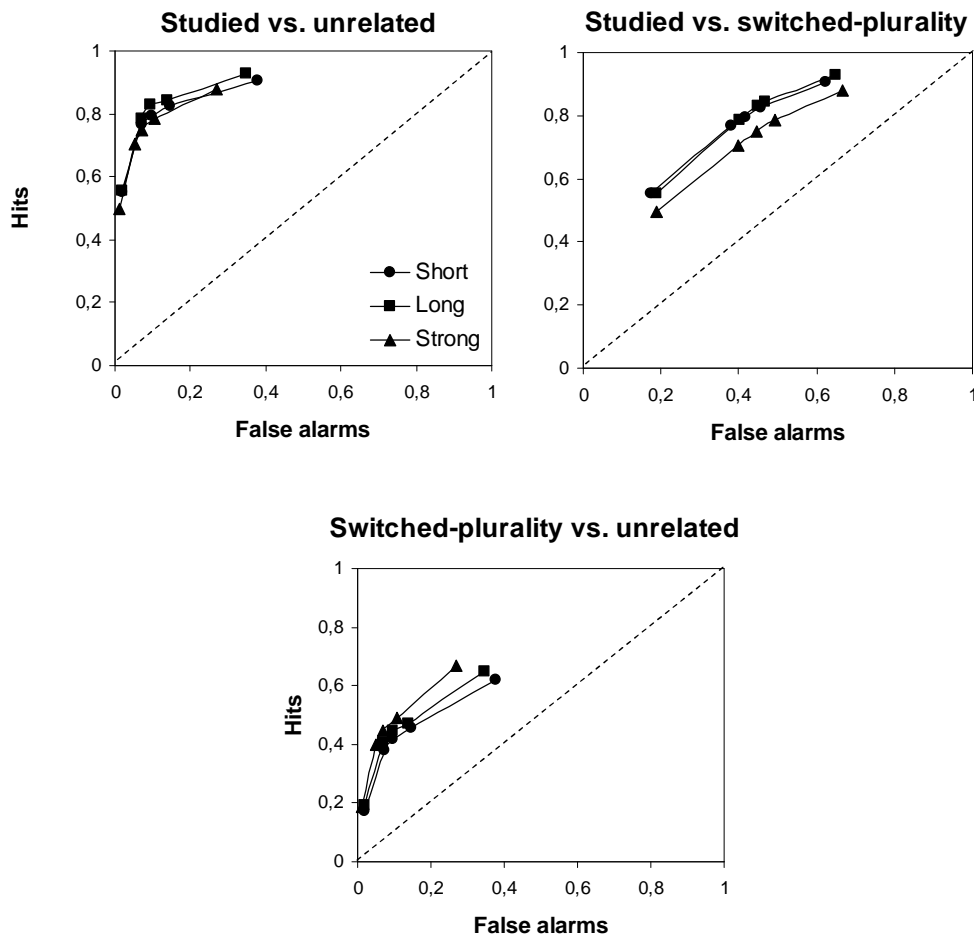
Separate 2 (encoding condition: *size* and *pleasantness*)  $\times$  3 (list type: *short*, *long* and *strong*) mixed-design ANOVAs on the discrimination measure ( $A_z$ ) were carried out for each discrimination type (SU, SSP, and SPU). There was no main effect of encoding condition and no interaction between list type and encoding condition for any of the discrimination types. Therefore, we analysed the data collapsed across encoding conditions.

**Table 3.7. Sensitivity ( $A_z$ ) across discrimination types (Exp. 3).**

List type	SU			SSP			SPU		
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>
<b>Short</b>	.91	$\tau$ $n$	.01	.74	$\tau$ $n$	.01	.72	$\tau$ $n$	.02
<b>Long</b>	.91	$\tau$ $\perp$	.01	.75	$\tau$ $\perp$	.01	.71	$\tau$ $\perp$	.02
<b>Strong</b>	.91	$n$ $\perp$	.01	.70	$**$ $\perp$	.01	.79	$***$ $\perp$	.01

*Note.*  $A_z$  = area under the ROC; SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality lures; SPU = switched-plurality vs. unrelated. *n* non-significant;  $** p < .01$ ;  $*** p < .001$ . Data collapsed across encoding conditions (size and pleasantness judgement).

Analysis of the aggregated  $A_z$  data showed no main effect of list type in the SU comparison,  $F < 1$ ,  $p = .44$ . In contrast, a main effect of list type was found in the SSP comparison,  $F(2, 234) = 7.08$ ,  $MSE = 0.01$ ,  $p < .01$ , such that sensitivity was lower for *strong* lists than for both *short* and *long* lists ( $ps < .01$ ; LSE) and such that sensitivity did not differ between *short* and *long* lists ( $p = .67$ ; no LLE). An effect of list type was also observed in the SPU comparison,  $F(2, 234) = 17.12$ ,  $MSE = 0.01$ ,  $p < .001$ , such that *pseudodiscrimination* was higher for *strong* lists than for both *short* and *long* lists ( $ps < .001$ ; negative LSE) and such that sensitivity did not differ between *short* and *long* lists ( $p = .71$ ; no LLE). Importantly, all results were significant in the *size* judgement condition if and only if they were significant in the *pleasantness* condition. Table 3.7 presents  $A_z$  data for each discrimination type collapsed across encoding tasks. Sensitivity ( $A_z$ ,  $d'$ ) and bias ( $c_a$ ,  $c$ ) for each encoding condition are listed in Appendix 1.



**Figure 3.8. ROC curves for Exp. 3.**

The ROC curves are superimposed in the “Studied vs. unrelated” comparison (non-significant difference). By contrast, in the “Studied vs. switched plurality” comparison, the curve for the *strong* list lies below the curves for both *short* and *long* lists ( $p = .001$ ) and in the “Switched plurality vs. unrelated” comparison, the *strong* curve lies above the other curves ( $p < .001$ ). Data collapsed across *size* and *pleasantness* encoding conditions. Pooled data ( $N = 119$ ).

We also carried out a 3 (list type: *short*, *long*, *strong*)  $\times$  2 (discrimination type: SU, SSP) repeated-measures ANOVA on  $A_z$  to investigate whether the impairment in sensitivity was specific to *strong* lists. A significant interaction was found,  $F(2, 234) = 10.56$ ,  $MSE = 0.01$ ,  $p < .001$ , confirming that there was no difference across list types for the SU comparison and that sensitivity was lower only for the *strong* list in the SSP comparison. Figure 3.8 illustrates these results. The ROC curve for *strong* lists lies below the curves for *short* and *long* lists in the SSP comparison and above them in the SPU comparison. By contrast, the curves largely coincide in the SU condition.

To rule out the possibility that interference from the other lists studied in the same experimental session could have masked the LLE, we conducted a 3 (list type)  $\times$  2 (discrimination type) mixed-design ANOVA on  $A_z$  only on lists that have been presented in the first study-test block (causing list type to become a between-participant manipulation). The results confirmed the pattern observed above: there was a significant interaction between list type and comparison type,  $F(2, 116) = 3.47$ ,  $MSE = 0.01$ ,  $p = .04$ , whereby an LSE was found only in the SSP comparison but no LLE was found in both SU and SSP comparisons. The interaction is significant despite the loss of power incurred by the change to a between-participant manipulation of list type.

To assess whether the null LLE in this experiment was a consequence of low statistical power, we carried out a power analysis on the list-length comparison (i.e., *short* vs. *long* lists). Using the  $d'$  values reported by Cary and Reder (2003, Exp. 3, SU comparison)<sup>5</sup> to compute the size of their LLE ( $g = 0.46$ ), we found that the estimated power to detect an LLE of that size in the SU comparison of our Experiment 3 ( $N = 132$ , two-tailed) was .99. In addition, the power to detect an LLE half the size of the effect found by Cary and Reder (2003, Exp. 3) was .75. Thus, the null effect observed here is unlikely to be a result of low power.

## Bias

Separate 2 (encoding condition: *size*, *pleasantness*)  $\times$  3 (list type: *short*, *long*, *strong*) mixed-design ANOVAs on the bias measure ( $c_a$ ) were conducted for each discrimination type (SU, SSP and SPU). For the SU comparison, the ANOVA revealed a main effect of list type,  $F(2,234) = 10.09$ ,  $MSE = 0.06$ ,  $p < .001$ , such that participants were more conservative in *strong* lists than in *short* and *long* lists ( $ps \leq .001$ ) and did not differ between *short* and *long* lists ( $p = .69$ ). There was no main effect of encoding task and no interaction. For the SSP comparison, there was no main effect of list type,  $F < 1$ ,  $p = .79$ ; there was however, a main effect of encoding condition,  $F(1,117) = 5.28$ ,  $MSE = 0.45$ , such that participants were more conservative in the *size* task than in the *pleasantness* task. There was also an

---

<sup>5</sup> The single-point sensitivity measure ( $d'$ ) was used here because it was the measure adopted by Cary and Reder (2003, Exp. 3) in their list-length study. Cary and Reder's (2003, Exp. 3) study was chosen because it was the experiment most similar to ours in which an LLE was found.

interaction between list type and encoding task,  $F(2,234) = 4.84$ ,  $MSE = 0.06$ ,  $p < .01$ , showing that participants were less conservative with *long* lists in the *size* condition but were equally conservative across lists in the *pleasantness* condition. Finally, for the SPU comparison, there was a marginal main effect of list type,  $F(2,234) = 10.09$ ,  $MSE = 0.06$ ,  $p = .08$ , such that participants were more conservative in *strong* lists. There was no main effect of encoding condition and no interaction. Table 3.8 presents these results collapsed across encoding tasks.

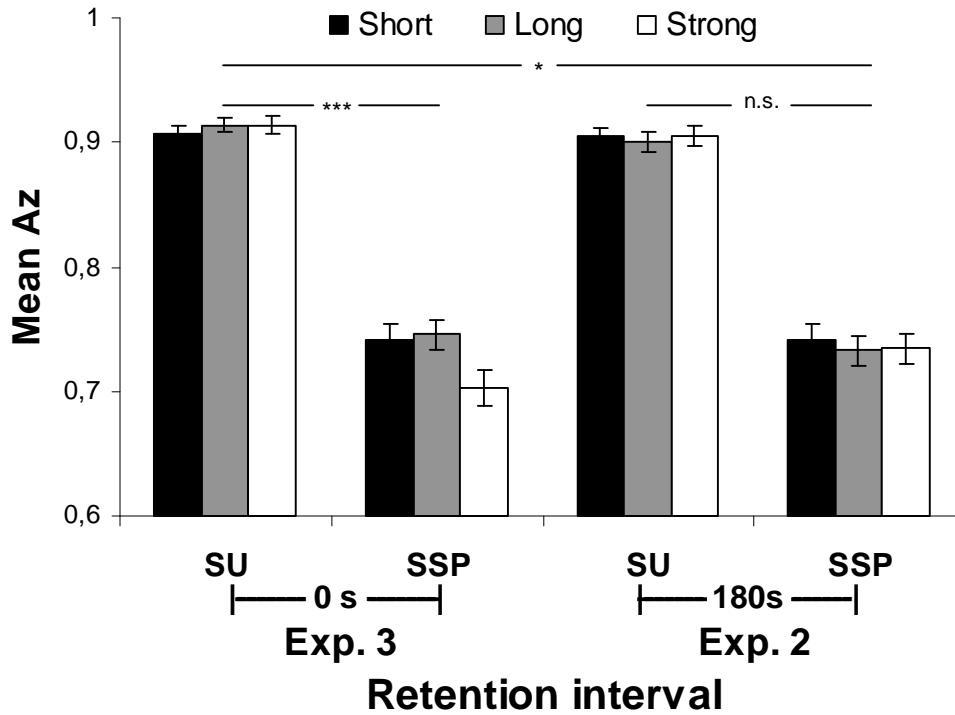
**Table 3.8. Bias ( $c_a$ ) across discrimination types (Exp. 3).**

List type	SU			SSP			SPU						
	$M$		$SEM$	$M$		$SEM$	$M$		$SEM$				
Short	0.19	$\top$ $n$	$\top$	0.03	-0.38	$\top$ $n$	$\top$	0.04	0.64	$\top$ $n$	$\top$	0.04	
Long	0.20	$\top$ ***	$\perp$	***	0.03	-0.39	$\top$ $n$	$\perp$	0.04	0.63	$\top$ $n$	$\perp$	0.04
Strong	0.32	$\perp$	$\perp$	0.03	-0.36	$n$ $\perp$	$\perp$	0.04	0.70	$*$ $\perp$	$\perp$	0.04	

*Note.*  $c_a$  = response bias (from  $X_3$ ); SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality lures; SPU = switched-plurality vs. unrelated.  $n$  non-significant;  $\dagger p < .10$ ;  $*$   $p < .05$ ;  $** p < .01$ ;  $*** p \leq .001$ . Data collapsed across encoding conditions (size and pleasantness judgement).

### Experiment 2 vs. Experiment 3 (Retention interval)

In Experiment 2, there was a 180-s retention interval for *long* and *strong* lists, whereas in Experiment 3 there was no retention interval. In the former, no LLE and no LSE were found, whereas in the latter an LSE was found in the SSP comparison. One natural question is to check whether the interaction between comparison type (SU vs. SSP) and retention interval (180 s in Exp. 2 vs. 0 s in Exp.3) is significant. This between-experiment comparison, however, needs to be conducted with two caveats. First, the experiments differ in more than one way. In Experiment 2, only lure type was manipulated, whereas in Experiment 3 both encoding task and lure type were manipulated. The added variable, however, did not interact with lure type. Thus, Experiment 3 may be treated as if it had a single encoding task. The second caveat is that participants in Experiment 3 were tested after all participants in Experiment 2 were tested. This temporal shift may introduce uncontrolled differences in the comparison between experiments (though we are not aware of any conspicuous time differences that could significantly affect the results).



**Figure 3.9. Sensitivity across retention intervals (Exps. 2 / 3).**

When retention interval is short (0 s for *long* and *strong* lists in Experiment 3), sensitivity in *strong* lists is impaired in SSP comparisons. When retention interval is long (180 s in Experiment 2), the differences between lists disappears. Significance values (\*, \*\*\*) refer to interaction terms between list type and comparison type. SU = studied vs. unrelated lure; SSP = studied vs. switched-plurality.  $A_z$  = sensitivity; *n.s.* non-significant; \*  $p < .05$ ; \*\*\*  $p < .001$ . Error bars = SEM.  $N = 233$ .

With those caveats in mind, we conducted a 2 [experiment: 2 (180 s), 3 (0 s)]  $\times$  3 (list type: *short*, *long* and *strong*)  $\times$  3 (comparison type: SU, SSP) mixed-design ANOVAs on sensitivity ( $A_z$ ). The three-way ANOVA revealed a significant interaction between experiment, list type and comparison type,  $F(2, 462) = 3.34$ ,  $MSE = 0.01$ . Separate 3 (list type: *short*, *long* and *strong*)  $\times$  2 (comparison type: SU, SSP) mixed-design ANOVAs on sensitivity ( $A_z$ ) carried out for each experiment showed that the interaction between comparison type and list type is highly significant when retention interval is short (0 s in Experiment 3,  $p < .001$ ) but non-significant when retention interval is long (180 s in Experiment 2,  $p = .75$ ). Figure 3.9 graphically describes those results.

### 3.4.3. Discussion

The results of Experiment 3 show that discrimination for weak items in *strong* lists is lower than discrimination for weak items in both *short* and *long* lists. This difference was observed in SSP comparisons but not in SU comparisons. In

addition, *pseudodiscrimination* (SPU comparison) was significantly higher for *strong* lists. Together, the results show that it was more difficult for participants to tell apart *targets* from *SP lures* in *strong* lists than in both *short* and *long* lists.

These results replicate the data reported by Norman (2002, Exp. 2). Moreover, the LSE was obtained despite our use of a weaker strength manipulation (3 repetitions as opposed to 6 repetitions) and despite our use of a longer average encoding time ( $\approx 1.8$  s, instead of 1.15 s used by Norman, 2002, to minimise rehearsal borrowing). The LSE was significant for each encoding condition. Thus, the results from Experiment 3 also show that it is possible to obtain a positive LSE with encoding tasks, such as the pleasantness task, which are not assumed to involve a great degree of trace overlap. This contrasts with encoding tasks previously used to obtain LSEs, such as the size task (Norman, 1999, 2002) and an emotion judgement task (e.g., decide whether a face is happy or angry, Norman et al., 2008), which presumably allow a greater degree of confusability.

That size judgement may result in more trace overlap is indirectly supported by the fact that participants showed more reluctance to say “old” for lists studied in the *size* condition than in the *pleasantness* condition. Encoding task, however, did not interact with list type. This suggests that the lack of interaction also observed in Experiment 1 was not simply a consequence of the longer retention interval adopted in that study.

The positive LSE in Experiment 3 contrasts with the null LSE in Experiment 2. A between-experiment ANOVA revealed a significant interaction between retention interval (between experiments) and comparison type (within experiment), confirming that discriminability for *strong* lists in SSP comparisons was selectively impaired when retention interval was reduced from 180 s to 0 s. This suggests that retention interval was a critical factor underlying the emergence of an LSE in this experiment.

Surprisingly, no LLE has been observed in both SU and SSP comparisons, despite the presence of a reliable LSE. This is the first time such dissociation has been observed. We defer the discussion of this issue to section 3.6. Suffice it to say,



however, that the controls suggested by Dennis and Humphreys (2001) naturally reduce the size of list-length effects. In that sense, it is not that surprising that an LLE has not been found. What is more surprising is that in the same experimental context, sharing the same experimental controls, an LSE has been found.

### 3.5. Experiment 4: Lure type, retention interval, 6x

In Experiments 1 to 3, no LLE has been found. One possible reason for the lack of an effect is that the length manipulation was not powerful enough. It is important that we are able to produce an LLE to show that the design used here is not subject to some unforeseen confound or methodological fault. Therefore, we set out to obtain an LLE in this experiment by increasing the long-to-short list-length ratio.

The ratio was increased from 2:1 (120:60 words) to 3.5:1 (210:60 words). This falls short of the 4:1 ratio used by Cary and Reder (2003, Exp. 3, 80:20 words). Yet it allows us to increase the number of presentations of strong items in *strong* lists from 3 to 6, equalising the number of repetitions adopted by Norman (2002). If an LLE is obtained with this new ratio, it would provide evidence that the LLE not only exists (contrary to the view proposed by Dennis and Humphreys, 2001) but also that it is stronger than previously thought (because it would be observed with a manipulation less powerful than the one used by Cary and Reder, 2003).

Retention interval was also manipulated in this experiment (between participants). Although the comparison between Experiments 2 and 3 provided evidence that retention interval modulates interference effects, it is methodologically more appropriate to obtain the effect in the context of the same experiment.

#### 3.5.1. Methods

##### Participants

One-hundred and eight University of Warwick students (40 males; age:  $M = 20.7$ ,  $SD = 4.1$ ) participated in the study: 54 took part in the short retention interval condition and 54 in the long interval condition. Participants were tested individually. Each session took 55 minutes and participants were paid £6.

## Materials

Stimuli were 450 imageable, concrete, familiar and medium-frequency nouns from the MRC Psycholinguistic Database: mean imageability = 5.71 out of 7, range = 5.02-6.52; mean concreteness = 5.77 out of 7, range = 5.00-6.48; mean familiarity = 5.09, range = 4.00-6.16; mean Kučera-Francis frequency = 16.63 occurrences per million, range = 0-99; mean word length = 5.47, range = 3-10.<sup>6</sup> The words were screened for semantic relatedness as in Experiment 1 (see 3.2.1). Thirty words were used as fillers and the remaining 420 words were assigned to 14 groups of 30 words, matched for word characteristics. Of the 14 word groups, 3 consisted of targets, 8 consisted of interference words and 3 consisted of unrelated lures. Distinct samples were produced for each participant.

## Design and Procedure

Figure 3.10 illustrates the experimental design. Design and Procedure were identical to Experiment 2 with two differences. First, participants were randomly assigned to one of two retention interval conditions. Second, *long* lists were longer and *strong* lists were stronger than in Experiment 2. Retention interval was either short (525 s for *short* lists and 0 s for *long* and *strong* lists) or long (705 s for *short* lists and 180 s for *long* and *strong* lists). The levels of retention interval were thus the same as the ones used in Experiments 1–3. The long-to-short list-length ratio increased from 2:1 in Experiment 2 to 3.5:1 in this experiment (*short*: 60 words; *long*: 210 words). Finally, the number of presentations of strong items increased from 3 in Experiment 2 to 6 in this experiment (*strong*: 60 different words, 210 study trials). Retention interval (*short* vs. *long*) was manipulated between participants (with 54 participants in each condition) and list type (*short*, *long*, *strong*) was manipulated within participants. As in the previous experiments, the test list consisted of 60 words (15 old items, 15 SP lures and 30 unrelated lures) and response was self-paced.

---

<sup>6</sup> Word properties were slightly different between Experiments 3 and 4 as more words were needed in the latter in order to increase the long-to-short list-length ratio.

List type	Short retention interval		
	Study	Distractor	Test
Short	[AB]	525 S	[tA, spA, unr]
Long	[AB] [CDEFG]		[tA, spA, unr]
Strong	[AB] [BBBBB]		[tA, spA, unr]

List type	Long retention interval		
	Study	Distractor	Test
Short	[AB]	705 S	[tA, spA, unr]
Long	[AB] [CDEFG]	180 S	[tA, spA, unr]
Strong	[AB] [BBBBB]	180 S	[tA, spA, unr]

**Figure 3.10. Design of Experiment 4.**

A size judgement task was used at study and retention interval was manipulated between participants. A-G = groups of 30 words; [X,Y] = word groups X and Y were merged and the order of the resulting list was randomised; tA = targets were half of the words from group A; spA = switched-plurality lures were the other half of the words from group A; unr = unrelated lures.

### 3.5.2. Results

#### Hits and false alarms

A 2 (word type: *target*, *SP lure*, *unrelated lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) mixed-design ANOVA on proportion of “old” responses revealed a large main effect of list type,  $F(2,212) = 16.11$ ,  $MSE = 0.02$ ,  $p < .001$ , such that the proportion of “old” responses in *strong* lists was lower compared to the proportion in *short* and *long* lists. There was also a marginal interaction between word type and list type,  $F(4,424) = 2.18$ ,  $MSE = 0.01$ ,  $p = .07$ , suggesting that that the proportion of “old” responses in *strong* lists was lower for *targets*, *SP lures* and *unrelated lures* compared to responses in *short* lists but that the proportion of “old” responses in *long* lists was higher for *SP lures*

and *unrelated lures* compared to responses in *short* lists. There was no main effect of retention interval, no interaction between retention interval and list type or word type and no three-way interaction among all three variables,  $F_s < 1$ ,  $p_s > .56$ . Because retention interval did not interact with any other variable, we collapsed retention interval in the following analyses.

Separate one-way ANOVAs across list type (*short*, *long*, *strong*) were carried out on hits (proportion of “old” responses to *targets*), SP false alarms (proportion of “old” responses to *SP lures*) and unrelated false alarms (proportion of “old” responses to *unrelated lures*). For hits, there was an effect of list type,  $F(2,214) = 7.52$ ,  $MSE = 0.01$ ,  $p = .001$ , such that hits were lower for *strong* lists than for *short* and *long* lists ( $p_s \leq .01$ ) and did not differ between *short* and *long* lists ( $p = .23$ ). For SP false alarms, there was also an effect of list type,  $F(2,214) = 5.22$ ,  $MSE = 0.02$ ,  $p < .01$ : SP false alarms in *strong* lists were lower than in *short* lists ( $p < .01$ ) and SP false alarms in *long* lists were marginally higher than in *short* lists ( $p = .05$ ). Finally, for unrelated false alarms, there was a large effect of list type,  $F(2,214) = 22.60$ ,  $MSE = 0.01$ ,  $p < .001$ , showing that unrelated false alarms in *strong* lists were lower than in *short* and *long* lists ( $p_s < .001$ ). Unrelated false alarms between *short* and *long* lists did not reliably differ ( $p = .35$ ).

**Table 3.9. Hits and false alarms (Exp. 4).**

List type	HR Targets				FAR SP lures				FAR Unrelated			
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>	
<b>Short</b>	.78	τ	τ	.01	.44	τ	τ	.02	.13	τ	τ	.01
		n				*				n		
<b>Long</b>	.76	τ	⊥	***	.48	τ	⊥	.02	.14	τ	⊥	***
		*				**				***		
<b>Strong</b>	.72	⊥	⊥	.01	.42	⊥	⊥	.02	.08	⊥	⊥	.01

*Note.* SP = switched plurality; *n* non-significant; \*  $p \leq .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .  $N = 108$ . Data collapsed across retention interval (*short* and *long*).

Hits and false alarms, collapsed across retention intervals, are presented in Table 3.9. Hits and false alarms broken down by retention intervals are presented in Appendix 1 together with single-point measures of sensitivity ( $d'$ ) and bias ( $c$ ).

There was no effect of length in the SU comparison, as the interaction between word type (*target*, *unrelated lure*) and list type (*short*, *long*) was not significant ( $p = .13$ ). However, there was an effect of length in the SSP comparison, as the interaction between word type (*target*, *SP lure*) and list type (*short*, *long*) was significant,  $F(1,106) = 6.39$ ,  $MSE = 0.02$ . The interaction showed that hits decreased from *short* to *long* lists and false alarms increased from *short* to *long* lists. By contrast, there was no effect of list strength in both SU and SSP comparisons: the interactions between word type (*target* vs. *SP lure* and *target* vs. *unrelated lure*) and list type (*short*, *strong*) were not significant ( $ps > .12$ ). The interaction was not significant because the decrease in hits from *short* to *strong* lists was not large enough to compensate for the concurrent decrease in SP false alarms and unrelated false alarms.

### Sensitivity

A total of 972 unequal-variance Gaussian models were fitted to individual participants' confidence data; 9 models were excluded due to poor fits. The results refer to the parameter estimates of the remaining 100 participants. Table 3.10 summarises the results collapsed across retention intervals.

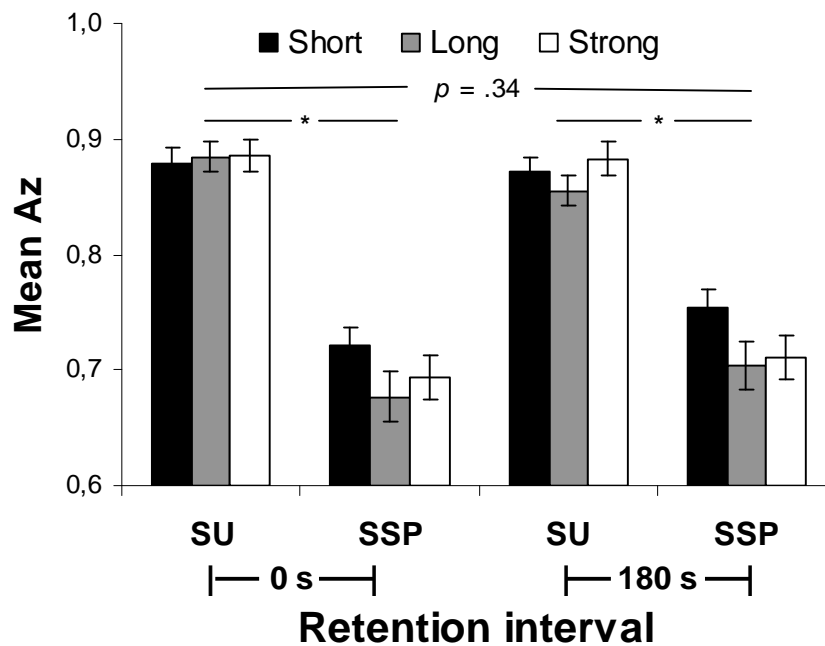
**Table 3.10. Sensitivity ( $A_z$ ) across discrimination types (Exp. 4).**

List type	SU				SSP				SPU			
	$M$			$SEM$	$M$			$SEM$	$M$			$SEM$
<b>Short</b>	.88	⊥	⊥	.01	.74	⊥	⊥	.01	.68	⊥	⊥	.02
		n	n			***	**			*		
<b>Long</b>	.87	⊥	⊥	.01	.69	⊥	⊥	.02	.72	⊥	⊥	.02
		⊥				⊥				⊥	***	
<b>Strong</b>	.88	⊥	⊥	.01	.70	⊥	⊥	.01	.75	⊥	⊥	.02
										*		

*Note.*  $A_z$  = area under the ROC; SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality lures; SPU = switched-plurality vs. unrelated.  $n$  non-significant; \*  $p \leq .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ . Data collapsed across retention interval (*short* and *long*).  $N = 100$ .

Separate 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) mixed-design ANOVAs were carried out on the sensitivity measure ( $A_z$ ) for each discrimination type (SU, SSP and SPU). There was no main effect of retention interval and list type and no interaction between the two variables in the SU discrimination,  $F_s < 1.33$ ,  $ps > .27$ . There was also no main effect of retention interval and no interaction between retention interval and list type in the SSP

comparison,  $F_s < 1.44$ ,  $p_s > .23$ . There was, however, a main effect of list type in the SSP comparison,  $F(2,196) = 6.36$ ,  $MSE = 0.01$ ,  $p < .01$ , showing that participants were worse at discriminating targets from lures in both *long* lists ( $p < .001$ ; LLE) and *strong* lists ( $p = .007$ ; LSE) compared to discrimination in *short* lists. There was no difference in discriminability between *long* and *strong* lists ( $p = .43$ ). Finally, for the SPU comparison, there was a main effect of list type,  $F(2,196) = 7.91$ ,  $MSE = 0.02$ ,  $p < .001$ , such that *pseudodiscriminability* was higher for *long* lists than for *short* lists ( $p = .03$ ; negative LLE), higher for *strong* lists than for *short* lists ( $p < .001$ ; negative LSE) and marginally higher for *strong* lists than for *long* lists ( $p = .05$ ). There was also an effect of retention interval in the SPU comparison,  $F(1,98) = 4.09$ ,  $MSE = 0.05$ , where *pseudodiscriminability* was higher in the *short interval* condition than in the *long interval* condition. Yet there was no interaction between retention interval and list type,  $F < 1$ ,  $p = .52$ .

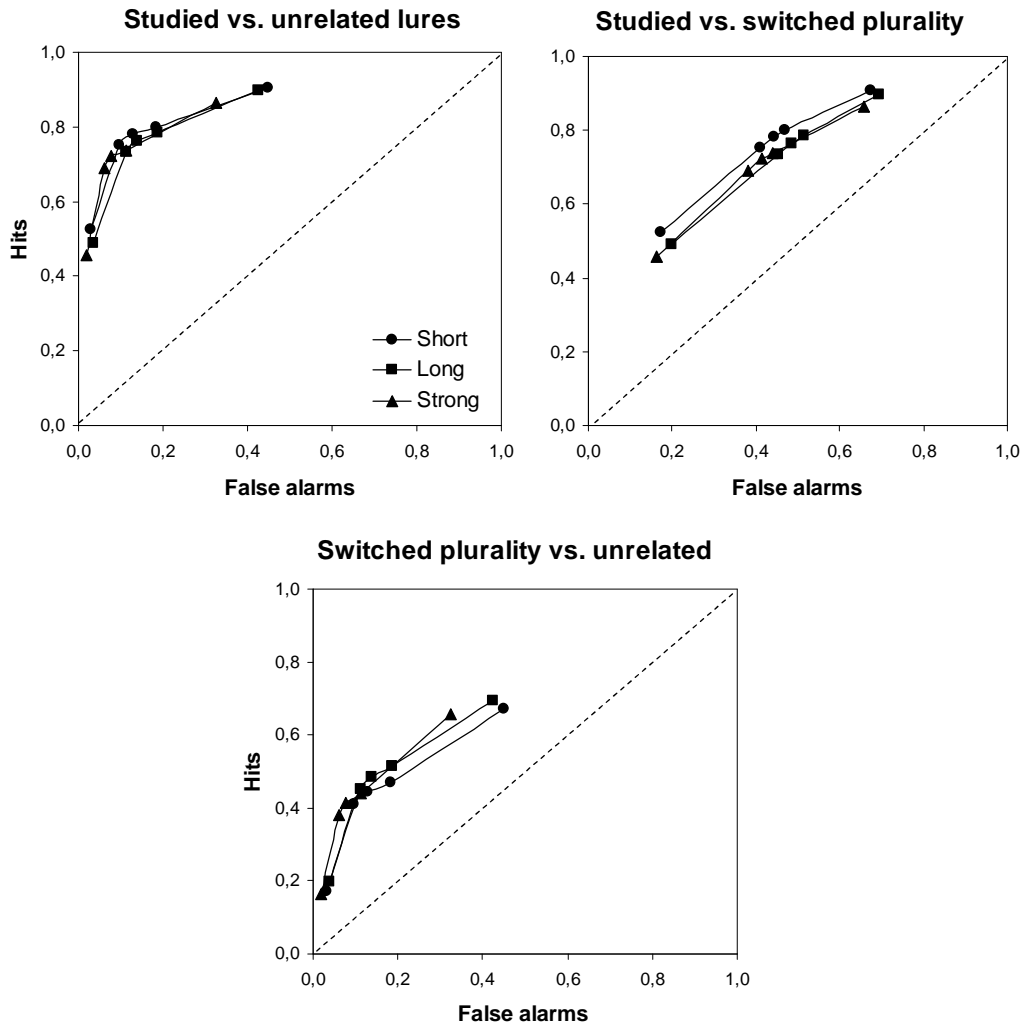


**Figure 3.11. Sensitivity across comparison types (Exp. 4).**

Sensitivity for *long* and *strong* lists was lower than for *short* lists in the SSP comparison but not in the SU comparison. The pattern was preserved across retention intervals. Significance values refer to interaction terms between list type and comparison type. SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality.  $A_z$  = sensitivity (area under ROC); \*  $p < .05$ . Error bars = SEM.

A 3 (list type: *short*, *long*, *strong*)  $\times$  2 (discrimination type: SU, SSP) repeated-measures ANOVA on  $A_z$  collapsed across retention intervals was conducted to investigate whether the impairment in sensitivity was specific to SSP comparisons. A significant interaction was found,  $F(2,196) = 7.57$ ,  $MSE = 0.01$ ,  $p$

= .001, confirming that there was no difference across list types for the SU comparison but that sensitivity was lower for *long* and *strong* lists in the SSP comparison in both short and long retention intervals. The three-way interaction between list type, discrimination type and retention interval was not significant,  $F(2,196) = 1.10, p = .34$ . Figure 3.11 illustrates these results.



**Figure 3.12. ROC curves across retention intervals (Exp. 4).**

ROC curves for *long* and *strong* lists lie below the curves for *short* lists in the “Studied vs. switched plurality” comparison and above them in the “Switched plurality vs. unrelated” comparison. These results illustrate list-length and list-strength effects. The same pattern occurs in both short and long retention intervals. By contrast, no effects were found in the “Studied vs. unrelated” comparison, where the curves largely overlapped.  $N = 100$ .

The differences in sensitivity across lists can also be observed in ROC curves.

Figure 3.12 shows the ROC curves, collapsed across retention intervals, for each list type and discrimination type. The ROC curves for *long* and *strong* lists lie below the curves for *short* lists in the SSP comparison and above them in the SPU

comparison (this is true for both short and long retention intervals). By contrast, the curves from the three list types largely overlap in the SU comparison.

To rule out the possibility that inter-list interference affected the results, a 3 (list type)  $\times$  2 (discrimination type) mixed-design ANOVA on  $A_z$  was carried out on data from the first study-test block only ( $N_{short} = 32$ ,  $N_{long} = 34$ ,  $N_{strong} = 34$ ). There was an interaction between list type and comparison type,  $F(2, 94) = 4.24$ ,  $MSE = 0.01$ , showing a trend towards a *negative* LSE in the SU comparison ( $M_{short} = .89$ ,  $M_{strong} = .91$ ,  $SEMs = .02$ ,  $p = .29$ ) and a marginally significant *positive* LSE in the SSP comparison ( $M_{short} = .75$ ,  $M_{strong} = .69$ ,  $SEMs = .02$ ,  $p = .07$ ). Negative LSEs in SU comparisons have been found elsewhere (Ratcliff et al., 1990). The LLEs were not significant, but performance for *long* lists was worse than for *short* lists in both comparisons (SU:  $M_{long} = .87$ ,  $SEM = .02$ ,  $p = .23$ ; SSP:  $M_{long} = .70$ ,  $SEM = .02$ ,  $p = .12$ ). Thus, the first-block analysis confirmed the results across blocks.

### Bias

Separate 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) mixed-design ANOVAs on the bias measure ( $c_a$ ) were carried out for each discrimination type (SU, SSP and SPU). For the SU comparison, the two-way ANOVA revealed a main effect of list type,  $F(2,196) = 30.94$ ,  $MSE = 0.06$ ,  $p < .001$ : participants were more conservative in *strong* lists than in *short* and *long* lists ( $ps < .001$ ) but similarly conservative in *short* and *long* lists ( $p = .14$ ). There was no main effect of retention interval and no interaction,  $F < 1.74$ ,  $p > .19$ . For the SSP comparison, there was also a main effect of list type,  $F(2,196) = 9.18$ ,  $MSE = 0.08$ ,  $p < .001$ : participants were more conservative in *strong* lists ( $ps \leq .001$ ) but equal in *short* and *long* lists ( $p = .73$ ).; there was no effect of retention interval and no interaction,  $F_s < 1.31$ ,  $ps > .27$ . For the SPU comparison, there was a main effect of list type,  $F(2,196) = 20.21$ ,  $MSE = 0.07$ ,  $p < .001$ , showing that participants were more conservative with *strong* lists ( $ps < .001$ ) but equal in *short* and *long* lists ( $p = .32$ ). There was no effect of retention interval,  $F < 1$ ,  $p = .61$ , but there was a marginal interaction between list type and retention interval,  $F(2,196) = 2.58$ ,  $MSE = 0.07$ ,  $p = .08$ , suggesting that the increase in bias in the SPU comparison was higher for *strong* lists when retention interval was short than



when the interval was long. Table 3.11 shows the results collapsed across intervals (data broken down by retention intervals is presented in Appendix 1).

**Table 3.11. Bias ( $c_a$ ) across discrimination types (Exp. 4).**

List type	SU				SSP				SPU			
	$M$			$SEM$	$M$			$SEM$	$M$			$SEM$
<b>Short</b>	0.16	$\top$	$\top$	0.03	-.30	$\top$	$\top$	0.04	0.60	$\top$	$\top$	0.04
		$n$				$n$				$n$		
<b>Long</b>	0.21	$\top$	$\perp$ ***	0.03	-.32	$\top$	$\perp$ ***	0.03	0.55	$\top$	$\perp$ ***	0.04
		***				***				***		
<b>Strong</b>	0.41	$\perp$	$\perp$	0.03	-.17	$\perp$	$\perp$	0.03	0.77	$\perp$	$\perp$	0.04

*Note.*  $c_a$  = response bias; SU = studied vs. unrelated; SSP = studied vs. switched plurality; SPU = switched plurality vs. unrelated.  $n$  non-significant; \*\*\*  $p \leq .001$ . Retention intervals were collapsed.

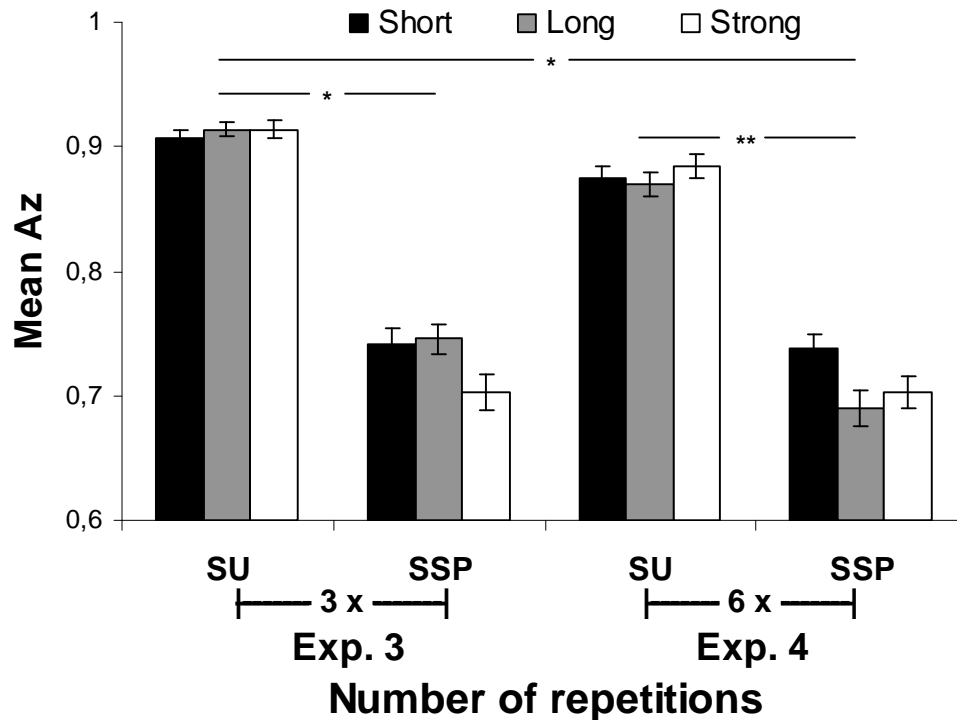
### Experiment 3 vs. Experiment 4 (number of repetitions)

Experiment 3 was carried out with weaker list-length and list-strength manipulations than Experiment 4. In the former, an LSE but no LLE was found in the SSP comparison, whereas in the latter both an LLE and an LSE were found in the SSP comparison. In order to confirm this pattern statistically, it is important to evaluate whether the interaction between comparison type (SU vs. SSP) and manipulation strength (3x in Exp. 3 vs. 6x in Exp.4) is significant.

As with the between-experiment comparison conducted in the last section (3.4.2), some issues must be borne in mind. First, Experiments 3 and 4 differ in several ways. In Experiment 3 encoding task was manipulated, whereas in Experiment 4 it was not. Conversely, in Experiment 4 retention interval was manipulated, whereas in Experiment 3 it was not. Both factors, however, can be safely ignored (i.e., analysed with their levels collapsed), since neither encoding task (in Experiment 3) nor retention interval (in Experiment 4) interacted with list type in analyses involving sensitivity. A second difference between the experiments is that participants in Experiment 4 were tested after all participants in Experiment 3. Although there is no obvious reason why such time difference would affect the results, it nonetheless introduces an additional source of variance.

There is, however, a third, and more important, difference between Experiments 3 and 4. Not only length ratio and number of repetitions were higher in Experiment 4 but also the average study-test lags. In a *short* list, for example, the delay

between study and test of a given target item was *at least* 210 s in Experiment 3; by contrast, this delay was at least 525 s (short interval) or 705 s (long interval) in Experiment 4. Any change in sensitivity between Experiments 3 and 4 could thus be caused either by the increase in manipulation strength or by the increase in study-test lag. Although study-test lag could affect discrimination independently of the length and strength changes across experiments, the data suggest that lag affected mostly SU but not SSP comparisons (which are the critical ones here).



**Figure 3.13. Sensitivity across number of repetitions (Exps. 3 / 4).**

Sensitivity for *strong* lists was lower than for *short* lists (LSE) in the SSP comparison when strong items were presented 3 and 6 times. Sensitivity for *long* lists was lower than for *short* lists (LLE) in the SSP comparison when long-to-short list-length ratio was 3.5:1 (Exp. 4) but not when it was 2:1 (Exp. 3). Neither effect was found in SU comparisons. Significance values (\*, \*\*) refer to interaction terms between list type and comparison type. SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality.  $A_z$  = sensitivity; \*  $p < .05$ ; \*\*  $p < .01$ . Error bars = SEM.  $N = 219$ .

The specific effect of study-test lag may be gauged by comparing sensitivity in *short* lists across experiments, as the only difference between *short* lists in Experiment 3 (apart from the encoding manipulation) and *short* lists in Experiment 4 was the retention interval (data across retention intervals within Experiment 4 were collapsed, since there was no difference between *short* and *long intervals*,  $t < 1$ ,  $p = .66$ ). Sensitivity for *short* lists in SU comparisons was higher in Experiment 3 ( $M = .91$ ,  $SEM = .01$ ) than in Experiment 4 ( $M = .88$ ,  $SEM = .01$ ),  $t(217) = 2.73$ ,  $p < .01$ , suggesting that the increase in retention interval

from 210 s to 525 s (or from 210 s to 705 s) had a negative impact in SU discrimination. By contrast, sensitivity for *short* lists in SSP comparisons did not change between experiments ( $M = .74$ ,  $SEM = .01$ ,  $t < 1$ ,  $p = .83$ ). A two-way ANOVA on sensitivity with study-test lag [210 s (Exp. 3) vs. 525 s/705 s (Exp. 4)] and comparison type (SU vs. SSP) as the independent variables confirmed those results: sensitivity for *short* lists decreased with longer study-test lags in SU comparisons but did not change in SSP comparisons [ $F(1,217) = 3.21$ ,  $MSE = 0.01$ ,  $p = .07$  (interaction term)]. Taken together, these results suggest that, although the change in study-test lag between Experiments 3 and 4 could cloud interpretations in SU comparisons across list types between experiments, it should not largely influence the results in SSP comparisons. Figure 3.13 illustrates the results across experiments, comparison types and list types.

A 2 [experiment: 3 (3x), 4 (6x)]  $\times$  3 (list type: *short*, *long*, *strong*)  $\times$  2 (comparison type: SU, SSP) mixed-design ANOVA on sensitivity  $A_z$  revealed a significant three-way interaction,  $F(2,434) = 3.67$ ,  $MSE = 0.01$ , suggesting that the interaction between list type and comparison type differed between experiments. Separate two-way ANOVAs for each experiment confirmed the trend: in Experiment 3, there was no difference across lists in the SU comparison and an LSE in the SSP comparison; in Experiment 4, there was no difference across lists in the SU comparison and both an LLE and an LSE in the SSP comparison,  $F_s > 7.6$ ,  $p_s \leq .001$ . We repeated the between-experiment analysis with data from Experiment 4 restricted to the short interval condition, so that data from both experiments came from a short interval condition. The results from the three-way ANOVA and the separate two-way ANOVAs did not change: there was a significant three-way interaction,  $F(2,332) = 4.69$ ,  $MSE = 0.01$  and significant two-way interactions,  $F_s > 4.76$ ,  $p_s \leq .01$ , between list type and comparison type.

### 3.5.3. Discussion

In Experiment 4, positive LLE and LSE were obtained in SSP comparisons and negative LLE and LSE were obtained SPU comparisons. By contrast, neither effect was found in SU comparisons. The LSE result is consistent with Norman (1999, 2002) and the LLE result is consistent with Cary and Reder (2003). The

data provide converging evidence that LLE is a real effect, contrary to the view put forward by Dennis and Humphreys (2001). The LLE was observed despite a series of effect-reducing controls and despite a manipulation less powerful than the one previously used to obtain the effect in similar conditions (Cary and Reder, 2003, Exp. 3). Note, however, that the LLE here was found in an SSP comparison, whereas in Cary and Reder (2003), it was found in an SU comparison.

The usefulness of the signal detection approach to data analysis becomes evident in this experiment. The analysis of hits and false alarms showed a harmful effect of the length manipulation on memory (hits fell and SP false alarms rose from *short* to *long* lists) but did not show any effect of the list strength manipulation (both hits and false alarms fell from *short* to *strong* lists but did not interact). This is in contrast to the sensitivity ( $A_z$ ) results, derived from SDT, which clearly show the presence of both LLE and LSE. The discrepancy is due to the inadequacy of the high-threshold-model assumption underlying the analysis of raw data (see 2.3.1). In fact, when data from the SSP comparison are fitted with a high-threshold model, 88 of the 324 models produced ( $108 \text{ participants} \times 3 \text{ list types}$ ) have to be rejected due to poor fits ( $\chi^2 p\text{-value} < .05$ ), 51 of which were very poor fits ( $\chi^2 p\text{-value} < .01$ ). By contrast, only 1 out of 324 unequal-variance SDT models did not fit the data well ( $\chi^2 p\text{-value} = .02$ ). Thus, we would have missed the reliable LSE produced in Experiment 4 if we had looked only at the raw data.

Unlike Experiments 2 and 3, where an interaction between list type and retention interval has been found (revealing an LSE only in the short interval condition), no such interaction was observed in this experiment: an LSE was found at both short and long retention intervals. It is possible that this lack of interaction was a result of a weak retention interval manipulation, since there was no main effect of retention interval on sensitivity in SU and SSP comparisons.

However, there is some indication that retention interval did affect memory in the expected manner. Response times, for example, indicate that recognition speed was the same across lists in the long interval condition but was faster to *short* lists than to *strong* lists in the short interval condition, hinting that discrimination in *strong* lists may have been more difficult when retention interval was short. Also,

pseudodiscrimination (SPU comparison on  $A_z$ ) increased from long to short retention intervals, suggesting that it was harder for participants to engage in recall-to-reject in the short interval condition.

Nonetheless, there was no interaction between retention interval and list type in any of the sensitivity comparisons (SU, SSP and SPU), indicating that retention interval did not differentially decrease  $A_z$  for *long* and *strong* lists when retention interval was short (as predicted by memory models such as CLS and BCDMEM).

One possible reason for the low retention interval effects is that the manipulation here was not sufficiently conspicuous relative to total study-test lag. The relative difference between short and long retention intervals was about half in Experiments 2 and 3 [ $\approx (180 \text{ s} - 0 \text{ s}) / 390 \text{ s}$ ] compared to a quarter in Experiment 4 [ $\approx (180 \text{ s} - 0 \text{ s}) / 705 \text{ s}$ ]. Thus, the difference between interval conditions was higher, and arguably of more consequence, in the former than in the latter case. If such interval ratio were indeed the reason behind the null retention interval effects, then one would expect to find an interaction between retention interval and list type on measures of bias ( $c_a$ ) in Experiments 2 and 3 but not in Experiment 4. That is because lists perceived to be more memorable tend to require more evidence for a positive response (high criterion setting) than lists perceived as less memorable (Hirshman, 1995; Singer & Wixted, 2006). Hence, bias should be more conservative during the test of lists learned relatively recently (short interval) compared to lists learned less recently (long interval). To the extent that this difference is more conspicuous in Experiments 2 and 3, a stronger shift in bias should be observed in those experiments than in Experiment 4.

The results, however, show the opposite pattern: there is no interaction between interval condition and list type in Experiments 2 and 3 for any comparison type (all  $ps > .35$ ), whereas there is a marginal interaction for the SPU comparison in Experiment 4 ( $p = .08$ ). The interaction showed that participants were more wary of responding “old” to *strong* lists in the short interval condition than in the long interval condition. This suggests that retention interval affected total list strength in *strong* lists sufficiently to alter the setting of decision criteria: participants treated *strong* lists in the short interval condition as if they were more memorable,

although this increase in total list strength was not enough to yield a larger LSE. Thus, it is unlikely that the absence of retention interval effects in this experiment compared to the effects in Experiments 2 and 3 can be accounted for by the relative difference between short and long retention intervals.

Another possible reason why retention interval did not modulate the interference effects here is that the magnitude of the effects may have peaked. Increasing list-strength may be effective in eliciting an effect up to a certain level, beyond which participants' idiosyncratic behaviours may progressively play a more important role. If the same item is repeated too many times, participants may decide they already know the item well and start to allocate less and less attention to that item in subsequent repetitions. In particular, participants may ignore repetitions of the item's plurality information. To the extent that this happens, it will reduce the effectiveness of increases in repetitions in the production of an LSE. Consistent with this view, participants spent less time on average studying repeated items in Experiment 4 [6 presentations;  $M = 1.55$  s, 95%  $CI$  (1.51 s, 1.58 s)] than in Experiment 3 [3 presentations;  $M = 1.69$  s, 95%  $CI$  (1.65 s, 1.73 s)]. Further indication that the LSE may have peaked (at least in the context of the current experimental design) comes from the fact that the effect sizes in Experiments 3 and 4 were very similar ( $d_z$ s = 0.42 and 0.39, respectively) and close to the effect size obtained by Norman (2002, Exp. 2,  $d_z = 0.44$ ). If the LSE has indeed peaked, as we believe it has under the present design, then attempts to modulate the effect through retention interval should prove unsuccessful.

The LLE results, on the other hand, hint at the possibility that the effect has some scope for improvement: the effect sizes on  $d'$  in Experiment 4 ( $d_z = 0.24$ , *SP lures*;  $d_z = 0.16$ , *unrelated lures*) were smaller than the effect size obtained from Cary and Reder's (2003, Exp. 3) data ( $g = 0.46$ ; *unrelated lures*). Yet it may not be easy to further increase the size of LLE with increasing list-length ratios. Participants may decide, after studying a certain number of words in a long list, that they will not be able to remember the items, possibly leading to lower task engagement. To the extent that this happens, performance differences between *short* and *long* lists (followed by complementary distractor tasks) would be reduced. Also, there is some evidence that retention interval may have little effect on LLEs: in Cary and

Reder (2003), when retention interval was increased from 0 s (Exp. 1) to 300 s (Exp. 2), effect sizes remained largely unchanged ( $g = 0.98$  vs.  $1.03$ , respectively). However, it is difficult to draw conclusions from those experiments as they included some confounds, namely, proactive design and variable-sized test lists. Moreover, study time also varied between their Experiments 1 and 2. Thus, the modulatory role of retention interval on LLE and LSE is still an open question.

In sum, Experiment 4 showed clear list-length and list-strength effects. Moreover, comparison of Experiments 3 and 4 indicated that, at least at lower manipulation strengths (3 presentations; length ratio = 2:1), an increase in the number of repetitions causes more interference than an increase in list-length ratio. At higher manipulation strengths (6 presentations; length ratio = 3.5:1), on the other hand, length interference appears to catch up, hindering performance to a similar degree as strength interference.

### **3.6. Discussion of Experiments 1 to 4**

#### **3.6.1. Empirical summary**

In Experiments 1 to 4 we aimed to test the hypotheses (*i*) that previous null LLE and LSE (Dennis & Humphreys, 2001) resulted from weak recollection at test, (*ii*) that previous results were not influenced by confounds in the experimental design (e.g., long encoding times, allowing for rehearsal borrowing), (*iii*) that the contrasting results found by Norman (2002) and Dennis and Humphreys (2001) were not caused by differences in encoding task, (*iv*) that retention interval modulates the size of LLE and LSE, and (*v*) that the strength of the manipulation (number of repetitions and list-length ratio) affects the size of the interference effects. The results provided partial support for the five hypotheses.

First, no LLE and no LSE were found in SU discrimination (assumed to involve less recollection) across all four experiments, replicating Dennis and Humphreys' (2001) null results. However, reliable LSEs were found in recollection-dependent SSP discrimination (Experiments 3 and 4) and SPU pseudodiscrimination

(Experiments 2, 3 and 4). The results indicate that Dennis and Humphreys' (2001) null LSE may have been caused by limited contribution of recollection at test.

Second, potential confounds identified in Dennis and Humphreys (2001; long encoding times) and Norman (2002; strong list being also longer) were unlikely to be critical to their results. In Experiments 1 to 3, no LLE was found, despite our use of short encoding times ( $\approx 1.8$  s). These null results suggest that Dennis and Humphreys' (2001) use of a long encoding time (3 s), which could lead participants to use some of the encoding time to rehearse previously studied items thereby reducing the LLE, was not the crucial factor behind their null result. Also, the fact that an LSE was found in Experiment 3 over and above any detrimental effect of list length, suggests that the LSE reported by Norman (2002) was not inflated by his comparison of a long, *strong* list with a short, *weak* list.

Third, encoding task did not have a significant effect on sensitivity measures (Experiments 1 and 3), although it did affect measures of bias and recognition speed, suggesting indirectly that the *size judgement* task entails more interference than the *pleasantness task*. Thus, our data provide no strong indication that effects of encoding task would be sufficiently large to explain the list-strength discrepancies between Dennis and Humphreys' (2001) and Norman's (2002) data, although the fact that we did not find such effects on sensitivity does not rule out the possibility that encoding task may systematically affect interference levels.

Fourth, retention interval appears to modulate LSE but not LLE. The results of Experiment 2 and 3 indicated that list-strength interference is larger when retention interval is short than when it is long. No such difference was observed in list-length manipulations. Moreover, the modulatory effect of retention interval was not replicated in Experiment 4. We address this issue in Experiments 5 to 7.

Finally, the magnitude of the interference manipulation affected LLE but not LSE in Experiment 4. An LLE was observed when the ratio between *long* and *short* list lengths was 3.5:1 but not when it was 2:1. By contrast, the same level of LSE was found regardless of the number of repetitions of the strong items in *strong* lists. The fact that an LLE was found in Experiment 4, despite our use of the controls



described by Dennis and Humphreys (2001), suggests that their null LLE might have resulted from a less-than-optimal manipulation (e.g., 2:1 list-length ratio). Although no difference in LSE size was observed across number of repetitions, it is possible that this null result was a consequence of LSE being at ceiling. We address this issue in Experiments 5 and 6.

### 3.6.2. Relation to other length and strength studies

The most common result in the literature has been the presence of LLE and the absence of LSE in recognition. How can the null LLE in Experiments 2 and 3 and the positive LSEs in Experiments 3 and 4 be reconciled with the published data?

The lack of a list-length effect in Experiments 2 and 3 can be accounted for by a combination of weak manipulation and the inclusion of controls known to reduce the effect size. Clearly, when the *long* list is not much longer than the *short* list, the likelihood of observing an LLE is diminished. In fact, few LLEs have been reported with length ratios as low as the ratio used in our Experiment 2 (2:1), and the results of these studies may have been marred by the confounds pointed out by Dennis and Humphreys (2001). For example, Zaki and Nosofsky (2001) reported data suggesting an LLE with a 2:1 length ratio (though they did not report the relevant inferential statistics, as they were pursuing another line of enquiry). The study, however, confounded study length with test length: longer study lists were followed by longer test lists. Thus, if the differences in discriminability between *short* and *long* lists were indeed significant, they could not be unambiguously attributed to the list-length manipulation. Similarly, Ratcliff et al. (1994, Exp. 3) reported an LLE between 16-word and 32-word lists. The fact that they used a proactive design, however, may have increased the size of the effect (e.g., due to differential loss of attention across lists). That proactive designs can increase the sizes of LLEs is attested by the experiments by Murnane and Shiffrin (1991a). In their study, length effects were higher when a proactive design was used (Exp. 1,  $d_z = 1.18$ ; Exp. 2,  $d_z = 1.29$ ; Exp. 4, equal arithmetic condition,  $d_z = 0.77$ ) than when a retroactive design was used (Exp. 3,  $d_z = 0.27$ ; Exp. 4, unequal arithmetic condition,  $d_z = 0.29$ ). Retroactive designs have been associated with LLEs only when accompanied by stronger manipulations. For instance, Gronlund and Elam

(1994, Exp. 1) reported a large LLE ( $d_z = 2.58$ ) using a retroactive design, although the effect may have been boosted by the use of a high length ratio (8:1).

Experiments using lists of categorised items avoid several of Dennis and Humphreys' (2001) confounds. The LLEs reported in those studies have also been obtained by means of larger length ratios (e.g., 2.5:1, Criss & Shiffrin, 2004a; 7:1, Ohrt & Gronlund, 1999; 3:1, Shiffrin et al., 1995). Moreover, the use of lists containing similar items increases confusability and, therefore, may add to the negative impact of list length on memory. By contrast, the studies reported here only used lists of unrelated items. Due to these differences, comparisons between studies using categorised and uncategorised stimuli may not be conclusive.

Our focus on the 2:1 list-length ratio is not based on a belief that this ratio has any special property. We focused on this ratio simply to make the point that, in agreement with Dennis and Humphreys (2001), LLEs in recognition are much harder to find than previously thought. When proper controls are in place, LLEs are only found with manipulations beyond a certain magnitude.

The controls suggested by Dennis and Humphreys (2001) and our weak manipulation naturally reduced the size of any existing LLE. In that sense, it is not that surprising that an LLE has not been found in Experiment 3. What is surprising is that, in the same experimental context, an LSE has been found. The presence of an LSE in Experiments 3 can be accounted for by a combination of high *target-lure* similarity and the use of stimuli that promote the use of recall-to-reject. Few reports exist in which an LSE has been found when only unrelated lures were used at test (Norman, 1999, Exp. 4a; 2002, Exp. 1). In those studies, strong items were presented 6 times. Surprisingly, these results were not fully replicated in a recent study where strong items were presented 11 times (Diana & Reder, 2005, Exp. 2)<sup>7</sup>. By contrast, most studies using only unrelated lures have failed to find LSEs (e.g., Hirshman, 1995; Murnane & Shiffrin, 1991a; Ratcliff et al., 1990; Ratcliff et al., 1992; Ratcliff et al., 1994; Yonelinas et al., 1992).

---

<sup>7</sup> Diana and Reder (2005) found an LSE when analysing *Remember* responses (*Remember* was lower for weak items in *strong* lists) but not when analysing overall hits and false alarms. Because they did not collect ROC data, it is not known to which extent the results are due to criterion shifts.

Most studies in which an LSE was found used similar *lures* at test (Norman, 2002; Norman et al., 2008; Verde & Rotello, 2004). Although important, the use of similar lures is not sufficient to elicit an LSE: Shiffrin, Huber, and Marinelli (1995) did use similar *lures* and yet found no LSE. To reconcile those results, one may argue that the lures they used (non-studied exemplars from studied categories) did not have a very similar counterpart in the study set in the way that the *SP lures* did, opening up the possibility that participants relied mostly on familiarity to discriminate between studied and non-studied items. This possibility is reinforced by the fact that participants in Shiffrin et al. (1995) were not explicitly encouraged to use recall at test (see Rotello et al., 2000, Exp. 2, for evidence that recall-to-reject is modulated by strategic control). Finally, Gallo (2004) provided evidence that participants will use recall-to-reject in categorised lists only when they are both instructed to do so and presented with categories that are short enough to allow the use of an exhaustive search strategy. Consistent with Gallo (2004), when category size in Shiffrin et al. (1995) was large (6 exemplars per category in Exp. 1), there was no trend towards an LSE, but when category size was small (2 exemplars per category in Exp. 2), there was a noticeable trend in  $d'$  towards an LSE (i.e., weak items in *mixed* lists were recognised less well than weak targets in *pure weak* lists; see Shiffrin et al., 1995, Figure 6). Thus, the use of similar *lures* coupled with instructions to use recall appears to boost LSEs.

### 3.6.3. Implications for memory models

Strength and length effects have been investigated in memory research partly because they can be used to test assumptions of computational models (for reviews, see Clark & Gronlund, 1996; Diana et al., 2006). The first theoretical question we addressed in Experiments 1 to 4 was whether LLE and LSE are more dependent on recollection than familiarity. The results of Experiments 3 and 4 clearly show that both LSE and (to a lesser extent) LLE are modulated by the engagement of recollection at test in the form of recall-to-reject.

These results support dual-process models, such as CLS and SAC, because they incorporate mechanisms that can account for the observed dissociations. In CLS,

when studied items are unrelated, strengthening some traces in memory (or adding new traces to memory) has the effect of reducing the weights and thus the activation of the other stored traces; the interference effect is more pronounced in the hippocampal model (recollection) than in the cortical model (familiarity) because in the former activation of lures is at floor whereas in the latter it is not and may consequently decrease with interference. In SAC, strengthening some items (or adding new items to memory) has the effect of reducing the activation of other items because there is less activation available to spread from context nodes, which are reactivated at test, to the episode nodes of the other studied items; the effect is more pronounced in episode nodes (recollection) because concept nodes (familiarity) have another source of activation, as they are also reinstated at test.<sup>8</sup>

The results of Experiments 3 and 4 do not support REM in that the model does not *a priori* predict an LSE (recall that REM was developed to account for null LSEs). The results presented here, however, are not strongly constraining on REM, since the model is flexible enough to fit positive, negative and null LSE results. The results, however, are constraining on BCDMEM: the model should be able to predict not only the presence of LLEs and LSEs but also their differential susceptibility to tests involving unrelated and highly similar lures. The effects observed here could not be easily attributed to confounds, such as contextual inertia, because the effects were also found in Experiment 4, where contextual inertia was presumably minimised in the long retention interval condition.

According to BCDMEM, only context noise (the number of contexts in which a word has been seen before) causes interference in recognition memory tasks, whereas item noise (the number and strength of words seen in the same context) should not matter. Because the present task is an item noise task, null LLE and LSE are predicted. The fact that we found positive effects challenges the context-noise assumption. However, that is not to say that context noise is irrelevant. Criss and Shiffrin (2004a, Exp. 2) showed that both the number of lists in which a target word appeared (context noise) and the number of words on the list that were similar to that target word (item noise) contributed to recognition. Thus, both item

---

<sup>8</sup> Note that in SAC interference occurs at retrieval, whereas in CLS it occurs at storage.

and context noises seem to underlie forgetting in recognition. Yet, the model, as presented in Dennis and Humphreys (2001), lacks an item-noise mechanism. Furthermore, it is unclear how BCDMEM could distinguish between targets (e.g., *banana*) and SP lures (e.g., *bananas*) and therefore account for the present results, since it represents words as individual nodes regardless of plurality.

Experiments 2, 3 and 4 also addressed the question of whether or not retention interval affects interference. Although most models would predict that stronger items in memory would entail stronger interference effects (i.e., that interference should be higher with short retention intervals), to our knowledge no direct test of this prediction has been carried out for both length and strength manipulations. Results confirming this prediction would support most models and fill an empirical gap; results disconfirming this prediction would present a problem for most models. Although Experiments 2 and 3 together suggest that retention interval may in fact modulate the size of LLE and LSE, Experiment 4 failed to replicate the same pattern. At this point we refrain from making strong claims about the retention interval results because they were observed in a comparison between experiments. We address this shortcoming in Chapter 4.

Another theoretical issue investigated here involves the impact of manipulation strength on the sizes of the interference effects. Although, from an empirical point of view, it may seem obvious that stronger manipulations should produce larger interference effects, from a theoretical perspective, that may not be the case. Both CLS and SAC predict a monotonic increase in LSE with interference strength, as repeatedly studying one item degrades the traces of other items (CLS) or reduces the activation spread to episode nodes (SAC). The fact that the LSE did not increase with extra strength, however, does not support the models' predictions. By contrast, REM predicts that LSE could disappear with more and more repetitions, since strong items would become so differentiated that they would contribute a negligible amount of activation at test, keeping sensitivity

unchanged.<sup>9</sup> The fact that LSEs were found in Experiments 3 and 4, regardless of manipulation strength, argues against REM's prediction of total differentiation.

The results also showed an increase in LLE in the SSP discrimination when the list-length ratio rose from 2:1 to 3.5:1, supporting CLS and SAC's predictions. Both CLS and SAC predict increases in LLE with longer lists, as adding new items degrades the traces of other items (CLS) or reduces the activation spread to the remaining episode nodes (SAC). REM, on the other hand, predicts little or no LLE when *lures* are too similar to *targets*, regardless of manipulation strength, since *SP lures* generate matches so strong that in effect they behave like *targets*; because *target* odds decrease with list length, so does *SP lure* odds, resulting in no difference, and no LLE in SSP discrimination (Criss & McClelland, 2006, Fig. 3).

Finally, comparison of Experiments 2 and 3 suggests that strength manipulations can cause more interference than length manipulations, at least at lower strengths (e.g., 3 presentations). If confirmed, the result may be problematic for SAC, which predicts larger length effects than strength effects, and for CLS, which predicts either equal-sized effects or larger effects of length. The results would also pose problems to BCDMEM to the extent that the model treats length and strength manipulations as equally irrelevant to interference in recognition.

#### 3.6.4. Limitations

The experiments in this chapter present several limitations that may reduce the generalisability of the results. First, average encoding times were shorter in *strong* lists than in *short* lists. Thus, even though retention interval was controlled across list types, study-test lag was not. Consequently, it is conceivable that an LSE was not found in SU comparisons in Experiments 1 to 4 because any decrease in discrimination associated with the strength manipulation could be compensated by an increase in discrimination due to the shorter study-test lags. To control for this, encoding times are kept constant in all the experiments in the next chapter.

---

<sup>9</sup> REM may predict larger LSEs with increasing strength if it is modified to include a recall-to-reject mechanism (Malmberg, Holden et al., 2004). However, it is not clear in that case how strengthening some items would reduce the recall of the other non-strengthened items.

Second, retention interval was manipulated between participants, thereby reducing experimental power. This may explain the lack of a modulatory effect of retention interval in Experiment 4. On the other hand, the comparison that did show an effect of retention interval was carried out between experiments (Experiments 2 and 3), which prevents us from drawing firm conclusions. To sidestep these issues, retention interval is manipulated within participants in the experiments presented in the next chapter.

Third, the retention interval manipulation may have inadvertently introduced qualitative differences between conditions. In the short retention interval condition, retention interval was reduced from 180 s, during which a distractor task was performed, to 0 s, where no distractor task was performed. Thus, the difference between retention intervals meant that in the long interval condition participants engaged in a video game task, whereas in the short interval condition they did not. This qualitative difference rather than the quantitative changes in retention intervals per se may underlie the effect in Experiments 2 and 3.

Finally, strong items were not tested in the experiments described in this chapter. Thus, it was not possible to assert directly whether strong items were indeed strengthened and to what extent. Although there was indirect evidence that the strength manipulation did work (e.g., change in bias in *strong* lists), it is important to confirm the magnitude of the strength manipulation by directly measuring sensitivity for strong items. Strong items are tested in Experiments 5 and 6.

To summarise, in the next chapter we describe four experiments that address the issues mentioned above. In particular, we attempt to replicate some of the effects obtained in Experiments 1 to 4 using an experimental design that is more similar to Norman (2002) and, arguably, more likely to produce interference effects.

## Chapter 4. Experiments 5-7

### 4.1. Introduction

In this chapter, we present four experiments designed to further test the boundary conditions of list-length and list-strength effects. As with the previous experiments, list type (*short, long, strong*) was manipulated within participants. Unlike the previous experiments, encoding time was fixed and short, strong items were also tested and retention interval was manipulated within participants. Moreover, short retention interval was set to 10 s rather than 0 s, to assess whether qualitative differences could explain the effects of interval observed in Experiments 2 and 3.

The design of this set of experiments closely followed Norman (2002, Exp. 2). In order to increase the likelihood of observing an LSE, Norman (2002) used a demanding encoding task (size judgement), a set of weakly related words at study and a short encoding time (1.15 s). Short encoding time coupled with a demanding encoding task should prevent participants from covertly rehearsing study items, which would work against finding the effects. Moreover, short encoding time coupled with a study list of unrelated words should reduce retrieval-practice effects (Anderson, 2003; see 5.2 for a discussion), given the strong dependency of those effects on semantic similarity (e.g., study words belonging to the same category).

Although the design adopted here incorporates the main features introduced by Norman (2002, Exp. 2), it departs from his design in that list-length and retention interval were also manipulated. Moreover, in Experiments 5 and 6, only related lures were present at test (i.e., SP lures). By using only related lures, we aimed to increase the contribution of recollection at test by forcing participants to rely on *recall-to-reject*, as familiarity alone is not diagnostic when targets and lures are so similar. Another advantage of using only related lures is that discriminability may increase compared to when unrelated lures are also tested (Heathcote, Raymond, & Dunn, 2006). Good discriminability is important because it facilitates the detection of differences among list types. Finally, by using only related lures we may be able to increase interference beyond its apparent peak level noted in Experiment 4.



In Experiments 5(*a, b*), strong items were presented 3 times, whereas in Experiment 6, strong items were presented 6 times. Because Experiment 6 contained a stronger manipulation, we also evaluated the impact of manipulation strength (number of repetitions and long-to-short length ratios) on the magnitude of LLE and LSE by comparing the effects across experiments. In Experiment 7, unrelated lures were presented at test in addition to SP lures. Because in Experiment 7 unrelated lures were also tested (*with-new* condition), we assessed whether the presence of unrelated lures can decrease overall discriminability by comparing those results with the results of Experiment 6 (*without-new* condition).

## 4.2. Experiment 5a: Retention interval, without new, 3x, one

In this experiment, we address some of the concerns raised in Chapter 3. First, strong items were tested together with weak items. This provides a direct measure of the success (or failure) of the strength manipulation. Second, only *targets* and *SP lures*, but not *unrelated lures*, were tested. In such test conditions, participants should rely more heavily on recollection than in conditions where *unrelated lures* are also tested. To the degree that LLE and LSE depend on recollection, higher recollection at test would aid the detection of higher levels of recollection-specific impairment. Consequently, we predict that the effect sizes in this experiment would be larger than the effect sizes in Experiments 2 and 3, where a similar manipulation magnitude was used (strong items were presented 3 times; list-length ratio was 2:1).

Retention interval was manipulated within participants. This raises the question of which presentation schedule is more appropriate: retention interval can either be manipulated in *one session* (i.e., participants take part in a sequence of study-test blocks for all list types and retention intervals within a single experimental session) or it can be manipulated in *two sessions* (i.e., participants experience one retention interval level per session and the sessions are conducted on different occasions).

Although apparently innocuous, the difference between one and two sessions may be both empirically and theoretically relevant. The difference is empirically relevant

because multi-list and single-list studies can yield different results. Gronlund and Elam (1994) found a dissociation between sensitivity ( $d'$ ) and estimated target variability ( $\sigma_T$ ) between multi-list and single-list sessions: they found an LLE but no change in target variability when multiple study-test blocks occurred during the same session (Exp. 1); when a single study-test block occurred in each session (Exp. 2), however, a larger LLE was found together with an increase in target variability. One interpretation is that inter-list variability may have swamped intra-list variability in their Experiment 1, also reducing the LLE's magnitude (percent changes in  $d'$  rose from 25% in Exp. 1 to 36% in Exp. 2). Consistent with this interpretation, when Gronlund and Elam (1994) separately analysed data from the first block of each session in Experiment 1, in effect simulating a single-list experiment, the result pattern paralleled the results of their Experiment 2. Thus, the number of study-test blocks may have consequences for interference effects.

The difference between one and two experimental sessions is also theoretically relevant because it enables the testing of an often-ignored assumption of global matching models, namely, that multi-list and single-list experiments should be equivalent. In SAM, lists can be isolated from each other through the setting of different context-to-item strengths (parameter  $a$  in the model). When items from the second study-test block are tested, items from the first block contribute little noise to the matching and decision processes because their strength ( $a$ ) to the current test context (which is associated with the second block) is small. Thus, according to SAM, results from multi-list and single-list experiments should not greatly differ. By contrast, Murdock and Kahana (1993a), in trying to account for the then accepted idea that LSEs did not arise in recognition, suggested that the noise added by other items in a study list should cause little interference at test given the noise of all other items previously studied by the participants (see 1.4.2). Consequently, according to Murdock and Kahana (1993a), it is immaterial whether participants take part in single-list or multi-list experiments: LSEs should not occur in either case. Gronlund and Elam's (1994) data suggest that neither assumption is entirely correct, since multi-list and single-list experiments can yield different results.

More importantly for our purposes, Gronlund and Elam's (1994) results suggest that the contribution of recollection in *short* lists in single-list studies may be more

prominent that in multi-list studies: not only the LLE increased from multi-list to single-list paradigm but also the ROC curves changed from symmetrical (Exp. 1, Fig. 1) to asymmetrical (Exp. 2, Fig. 4) as the leftmost point in the latter ROC moved up. Asymmetrical ROCs are interpreted, within the framework of dual-process models (e.g., Yonelinas, 1994), as evidence for recollection, since recalling an item increases the use of high-confidence “old” responses, represented as the leftmost point in ROC curves<sup>1</sup>. In *long* lists, on the other hand, the ROC were symmetrical in both single-list and multi-list tasks, indicating less recollection.

Because in our studies we hypothesise, as Norman (2002) did, that recollection is selectively affected by length and strength interference, it is important to maximise its role at test if we are to observe any changes in discriminability across list types. The results above thus suggest that splitting the retention interval manipulation into two sessions (two sets of 3 study-test blocks) may be a more efficient way of eliciting recollection (and, consequently, of observing interference effects) than manipulating retention interval in one session (one set of 6 study-test blocks). To test this hypothesis, we carried out two experiments with different retention interval schedules. In Experiment 5*a*, participants studied and were tested on all lists from both retention intervals in *one* session. In Experiment 5*b*, participants studied and were tested on lists from each retention interval in *two* sessions on different days. If the global matching models’ assumptions are correct (i.e., that lists can be perfectly isolated from each other or that lists add little noise to the memory system), then there should be little difference between the results of Experiments 5*a* and 5*b*. If, on the other hand, single-list and multi-list schedules differently affect recollection, as suggested by Gronlund and Elam’s (1994) results, then interference effects should be found in Experiment 5*b*, where recollection is more likely to be elicited, but not in Experiment 5*a*, where recollection is less likely.

#### 4.2.1. Methods

##### Participants

Forty-eight University of Warwick students (21 males; age:  $M = 21.5$ ,  $SD = 3.3$ ) were tested individually and paid £6 to take part in the study.

---

<sup>1</sup> Asymmetrical ROCs cannot be taken as unambiguous evidence for recollection, however, as single-process, unequal-variance SDT models also fit well asymmetrical ROCs (Heathcote et al., 2006).

## Materials

Stimuli consisted of 264 nouns from the MRC Psycholinguistic Database (mean imageability = 5.73; concreteness = 5.80; familiarity = 5.14; frequency = 17.6 occurrences per million; length = 5.48). Items were not strongly semantically related to one another. Words were classified as *targets* (*A* or *B*; if presented at study and at test in the same plurality), *SP lures* (*A* or *B*; if presented at study and at test with their plurality reversed) or *interference* (if presented at study but not at test). *SP lures* were constructed from half of the *targets* by reversing their plurality. Twenty-four words were fillers. The other 240 words were assigned to 8 matched groups of 30 words: 6 groups consisted of *targets/SP lures* (3 groups of *A items*; 3 groups of *B items*) and 2 groups consisted of *interference* items (groups *C* and *D* in Figure 4.1).

## Design

Each participant attended one 60-min. session. Each session consisted of six experimental blocks, each containing one of three different list types: *short* lists (30 *A* items and 30 *B* items presented once), *long* lists (30 *A* items and 30 *B* items presented once followed by 60 extra items) and *strong* lists (30 *A* items presented once and 30 *B* item presented 3 times). Participants were tested on all three list types, with list order balanced across participants. Retention interval was manipulated within-participants with order of short and long intervals counterbalanced between participants.

All the *targets* and *SP lures* used in the test phase originated from groups *A* and *B*. Half of the items in groups *A* and *B* were studied in their singular form; half were studied in their plural form. Items' plurals were generated by adding *s* to their singular form. *SP lures* were created by randomly sampling half of the items in each group (*A* or *B*) and reversing their plurality (singular to plural or vice-versa). This procedure was repeated for the same participant (so that the assignment of words to word groups and list types was different for short and long retention interval conditions) and repeated anew for each participant (so that the assignment of words to word groups and list types were also randomised across participants). Note that for *strong* lists, *B* items were subsequently repeated in the study phase, whereas for *short* and *long* lists, *B* items were presented only once. All target words were

presented before any of the interference items were repeated. *A* and *B* items were randomly intermixed at the beginning of each list. Thus, from the participants' perspective, *A* and *B* items were indistinguishable from each other.<sup>2</sup> Figure 4.1 illustrates this experimental design.

List type	Short retention interval		
	Study	Distractor	Test
Short	[AB]	109 s	[tA, spA] [tB, spB]
Long	[AB] [CD]	10 s	[tA, spA] [tB, spB]
Strong	[AB] [BB]	10 s	[tA, spA] [tB, spB]

List type	Long retention interval		
	Study	Distractor	Test
Short	[AB]	219 s	[tA, spA] [tB, spB]
Long	[AB] [CD]	120 s	[tA, spA] [tB, spB]
Strong	[AB] [BB]	120 s	[tA, spA] [tB, spB]

**Figure 4.1. Design of Experiment 5.**

A-D = matched groups of 30 words; [X,Y] = first, word groups X and Y are merged; then word order in the new list is randomised; tA = targets from group A; spA = switched-plurality lures from group A; tB = targets from group B; spB = switched-plurality lures from group B.

### Procedure

Stimuli were presented on a PC screen. Each experimental session consisted of seven blocks: one practice block and six experimental blocks. Each block consisted of three phases: study, distractor and test.

<sup>2</sup> Note that for *strong* lists, *B* items will become distinct from the participants' perspective when they begin to repeat. More importantly, however, is that they are not distinguishable from *A* items at the beginning of the list. This helps avoid participants' use of strategies, such as rehearsal borrowing.

*Study Phase.* In each study phase, participants were presented with either 60-word lists (*short*) or 120-word lists (*long* and *strong*). Four extra items were used as fillers (2 at the start and 2 at the end) of each study list to control for primacy and recency effects. Participants were warned that some items might appear several times. They were also informed that their memory would be tested. Participants were shown a shoebox (15 cm wide, 28 cm long, 10 cm deep) before the start of the experiment and were instructed to decide whether or not a typical instance of the word fits in the shoebox by pressing, respectively, a left (“yes”) or a right (“no”) button on a gamepad. They were told to pay attention to the plurality of the words. If the item was in its plural form, they should imagine two instances of that item in the box. Each item was displayed for 1,150 ms, with 500 ms of inter-stimulus interval.

*Distractor Phase.* A video game task was used to equate study-test lag across list types. Retention interval was manipulated by varying the duration of the video game task. In the long retention interval condition, the game lasted 219 s for *short* lists and 120 s for *long* and *strong* lists, whereas in the short retention interval condition, the game lasted 109 s for *short* lists and 10 s for *long* and *strong* lists.<sup>3</sup>

*Test Phase.* The test lists contained 60 words (15 *targets* randomly intermixed with 15 *SP lures*, both made up of *A* items, followed by 15 *targets* randomly intermixed with 15 *SP lures* from group *B*). Words appeared one at a time on the screen and response was self-paced. Subjects were instructed to rate their memory confidence on a scale from 1 to 6 (*definitely old* to *definitely new*). Participants were allowed to take a short break at the end of each block.

---

<sup>3</sup> The use of retention intervals of the order of 2 min for *long* and *strong* lists is justified by the properties of LTP (long-term potentiation), a phenomenon interpreted as the biological mechanism behind memory formation in the hippocampus. In the CLS model, LTP is the process underlying trace storage (learning) and interference (forgetting). *In vivo* and *in vitro* studies showed that, after a learning episode (i.e., LTP induction via electrical stimulation), neuronal activity in the stimulated area immediately rises, quickly decays during the first minutes and then slowly settles on a level of activity that is higher than before learning (Bliss & Collingridge, 1993). In our studies, we compare interference effects between 10 s and 120 s after the end of the study list because, in this time range, LTP should vary from high to low levels of activity. It is known that a 120-s interval does not prevent an LSE (Norman, 2002); our interest here is in observing a larger LSE in the short interval condition. Ideally, however, retention intervals would have to vary on a much wider range (from seconds to hours) to allow a more thorough testing of the LTP hypothesis of memory interference.

### 4.2.2. Results

#### Hits and false alarms: A items

A 2 (word type: *target*, *SP lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) within-participants ANOVA on proportion of “old” responses revealed a strong main effect of word type,  $F(1,47) = 232.56$ ,  $MSE = 0.01$ ,  $p < .001$ , such that the proportion of “old” responses was higher for *targets* ( $M = .69$ ,  $SEM = 0.01$ ) than for *SP lures* ( $M = .32$ ,  $SEM = 0.02$ ). The result shows that recognition was above chance (a potential concern, given the high target-lure similarity). All other main effects and interactions were not significant (all  $ps > .10$ ).

Separate two-way repeated-measures ANOVAs were carried out on the proportion of “old” responses for each word type (*target* and *SP lure*) with retention interval (*short* and *long*) and list type (*short*, *long* and *strong*) as the independent variables. For *targets*, there was a marginal effect of list type,  $F(2,94) = 2.99$ ,  $MSE = 0.02$ ,  $p = .06$ , such that the hit rates were lower for *strong* lists. There was no main effect of retention interval and no interaction ( $ps > .30$ ). For *SP lures*, there was no main effect of list type and no interaction between list type and retention interval ( $ps > .38$ ). There was, however, a main effect of retention interval,  $F(1,47) = 5.04$ ,  $MSE = 0.02$ , such that false alarms were lower when retention interval was short compared to when it was long. Hits and false alarms are presented in Table 4.1. Sensitivity ( $d'$ ) and bias ( $c$ ) for each retention interval are reported in Appendix 1.

There was no effect of list length, as the interaction between word type (*target* vs. *SP lure*) and list type (*short* vs. *long*) was not significant,  $F(1,47) = 1.67$ ,  $p = .20$ . There was also no effect of list strength, as the interaction between word type and list type (*short* vs. *strong*) was not significant,  $F(1,47) = 1.06$ ,  $p = .31$ . The interactions were not significant because both hits and false alarms decreased in *long* and *strong* lists and to a similar degree when compared to *short* lists (effective list length manipulations cause mirror effects, whereby hits decrease and false alarms increase in *long* lists relative to *short* lists; list strength manipulations, on the other hand, usually cause larger falls in hits in *strong* lists relative to *short* lists).

**Table 4.1. Hits and false alarms across item types (Exp. 5a).**

<b>A items</b>							
<b>List type</b>	<b>HR Targets</b>			<b>FAR SP lures</b>			
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	
<b>Short</b>	.72	$\tau$ <i>n</i>	$\tau$ *	.02	.34	$\tau$ <i>n</i>	$\tau$ <i>n</i>
<b>Long</b>	.69	$\tau$ <i>n</i>	$\perp$	.02	.32	$\tau$ <i>n</i>	$\perp$
<b>Strong</b>	.67	$\perp$	$\perp$	.02	.32	$\perp$	$\perp$
<b>B items</b>							
<b>List type</b>	<b>HR Targets</b>			<b>FAR SP lures</b>			
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	
<b>Short</b>	.70	$\tau$ <i>n</i>	$\tau$	.02	.33	$\tau$ <i>n</i>	$\tau$ <i>n</i>
<b>Long</b>	.71	$\tau$ ***	$\perp$	.02	.33	$\tau$ <i>n</i>	$\perp$
<b>Strong</b>	.83	$\perp$	$\perp$	.02	.30	$\perp$	$\perp$

*Note.* HR = hits; FAR = false alarms; SP = switched plurality; A/B items = items from the beginning of the study list (in *strong* lists, *B* items are repeated). *n* non-significant; \*  $p < .05$ ; \*\*\*  $p < .001$ . Data collapsed across short (10 s) and long (120 s) retention intervals.  $N = 48$ .

#### Hits and false alarms: *B* items

A 2 (word type: *target*, *SP lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) within-participants ANOVA on proportion of “old” responses revealed main effects of word type (more “old” responses to *targets*) and list type (more “old” responses to *strong* lists) and an interaction between the two variables (more “old” responses to *targets* and fewer “old” responses to *SP lures* in *strong* lists than in *short* and *long* lists;  $F_s > 7.7$ ,  $p_s \leq .001$ ). The results reflect the fact that *B* items were repeated in *strong* lists and indicate a *between-list strength-based mirror effect* (Stretch & Wixted, 1998b), whereby the proportion of “old” responses in *strong* lists increased for *targets* and decreased for *SP lures* compared to the proportion of “old” responses in *short* and *long* lists.

To confirm these results, separate two-way repeated-measures ANOVAs on the proportion of “old” responses were carried out for each word type (*target* and *SP lure*) with retention interval (*short* and *long*) and list type (*short*, *long* and *strong*) as the independent variables. For *targets*, there was a strong main effect of list type,  $F(2,94) = 25.46$ ,  $MSE = 0.02$ ,  $p < .001$ , such that there were more hits in *strong* lists,



reflecting better memory for repeated *B* items. There was no effect of retention interval and no interaction ( $ps > .50$ ). For *SP lures*, there was no main effect of list type and no interaction with retention interval ( $ps > .28$ ), although there was a trend for false alarms to be higher in the long retention interval condition,  $F(1,47) = 2.69$ ,  $MSE = 0.02$ ,  $p = .11$ . The latter results shows that the between-list strength-based mirror effect, suggested by the interaction between word type and list type in the three-way ANOVA above, was not complete: hits in *strong* lists increased but false alarms did not reliably decrease relative to the other list types. Hits and false alarms are presented in Table 4.1. Sensitivity ( $d'$ ) and bias ( $c$ ) are reported in Appendix 1.

#### Hits and false alarms: *A* and *B* (within-list strength-based effects)

To assess whether the strength manipulation yielded a *within-list strength-based mirror effect* (Stretch & Wixted, 1998b), we conducted a 2 (word type: *target*, *SP lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  2 (item strength: *A*, *B*) ANOVA on proportion of “old” only for *strong* lists. There was an interaction between word type and item strength such that proportion “old” for *targets* increased and for *SP lures* decreased from *A* (weak) to *B* (strong) items,  $F(1,47) = 41.23$ ,  $MSE = 0.02$ ,  $p < .001$ . The ANOVA also revealed an interaction between word type and retention interval, suggesting that the difference between hits and false alarms was larger when retention interval was short,  $F(1,47) = 5.03$ ,  $MSE = 0.02$ .

To confirm whether a mirror effect did in fact occur, two two-way ANOVAs on proportion “old” were conducted, one for each word type (*targets* and *SP lures*), with item strength and retention interval as the independent variables. The ANOVAs revealed that hits increased from *A* (weak) to *B* (strong) items ( $M_A = .67$ ,  $SEM = .02$ ;  $M_B = .83$ ,  $SEM = .01$ ;  $p < .001$ ) but that false alarms did not decrease ( $M_A = .32$ ,  $SEM = .02$ ;  $M_B = .30$ ,  $SEM = .03$ ;  $p = .40$ ). The result replicates the pattern observed in previous studies (Stretch & Wixted, 1998b; Verde & Rotello, 2007) and suggests that participants are reluctant to change their criterion within lists, even when there are clear boundaries between weak and strong items [weak items (*A targets* and *A SP lures*) were presented in the first half of the test list, whereas strong items (*B targets* and *B SP lures*) appeared in the second half].

### Sensitivity: *A* and *B* items

$A_z$  was estimated by fitting Gaussian models to confidence data. None of the 288 models [48 participants  $\times$  2 retention intervals (*short*, *long*)  $\times$  3 list types (*short*, *long* and *strong*)] for *A* items was rejected at the .05 level.

A 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) ANOVA conducted on  $A_z$  revealed a main effect of retention interval,  $F(1,47) = 6.60$ ,  $MSE = 0.01$ , showing that participants were overall better at discriminating targets from lures when retention interval was short ( $M = .76$ ,  $SEM = .02$ ) than when it was long ( $M = .72$ ,  $SEM = .02$ ). There was no effect of list type (i.e., no LLE and no LSE) and no interaction ( $F_s < 1$ ,  $p_s > .69$ ). The effect of retention interval indicates that the retention interval manipulation was effective.

**Table 4.2. Sensitivity ( $A_z$ ; SSP comparison) across item types (Exp. 5a).**

List type	A items			B items		
	<i>M</i>	<i>(N = 48)</i>		<i>M</i>	<i>(N = 46)</i>	
			<i>SEM</i>			
<b>Short</b>	.74	τ	τ	.74	τ	τ
		n	n		n	
<b>Long</b>	.75	τ	⊥	.75	τ	⊥
		n			***	***
<b>Strong</b>	.73	⊥	⊥	.83	⊥	⊥

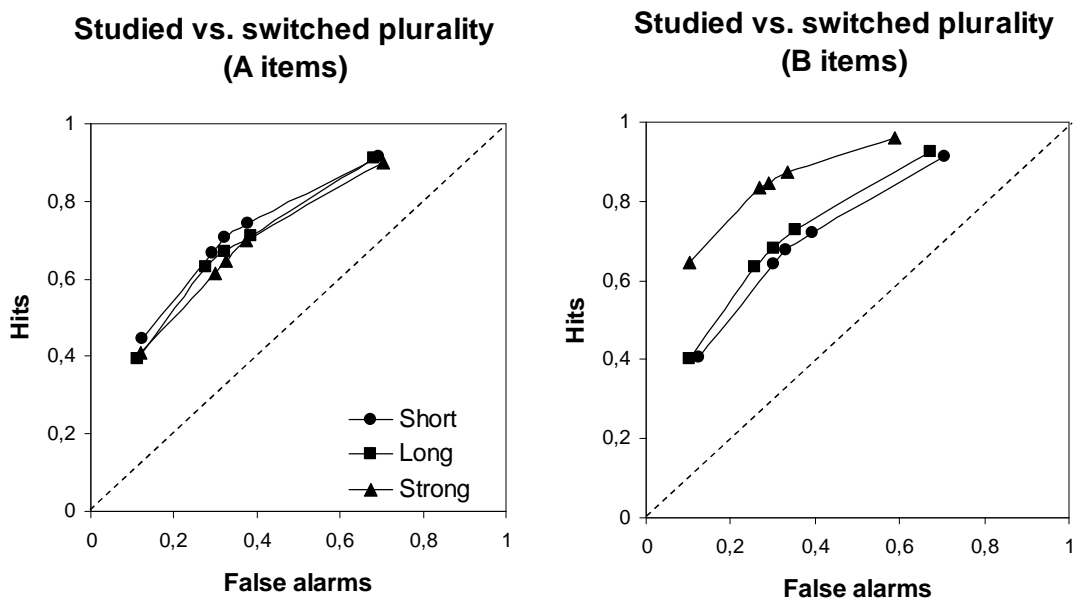
*Note.*  $A_z$  = area under the ROC; A/B items = items from the beginning of the study list (in *strong* lists, *B* items are repeated). *n* non-significant; \*\*\*  $p < .001$ . Data collapsed across retention intervals (short, 10 s, and long, 120 s).

For *B* items, two of the 288 unequal-variance SDT models fitted to the data were rejected at the .05 level. Results below refer to the data from the remaining 46 participants. Table 4.2 shows the results collapsed across retention intervals.

A 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) ANOVA conducted on  $A_z$  revealed a main effect of list type,  $F(2,90) = 25.38$ ,  $MSE = 0.01$ ,  $p < .001$ , showing that discriminability for *B* items was better for *strong* than for *short* and *long* lists (*B* items were presented three times in *strong* lists). The effect indicates that the strength manipulation was effective. There was no effect of retention interval and no interaction ( $F_s < 1.9$ ,  $p_s > .15$ ).

Because *B* items were indistinguishable from *A* items in *long* lists, they too should suffer interference. In other words, the experimental design was such that there was

in effect a list-length manipulation for *B* items as well. An LLE for *B* items would be found if discrimination for those items in *long* lists was lower than in *short* lists. A 2 (retention interval: *short*, *long*)  $\times$  2 (list type: *short*, *long*) on *B* items showed no main effects and no interactions ( $F_s < 1$ ,  $p_s > .50$ ), replicating the null LLE found with *A* items. When item type (*A* and *B*) was entered as an additional factor in the ANOVA, the result was the same: no effect of list type ( $F < 1$ ,  $p = .57$ ). There was, however, a marginal interaction between item strength and retention interval, reflecting the fact that retention interval affected sensitivity for *A* items but not for *B* items,  $F(1,47) = 3.39$ ,  $MSE = 0.01$ ,  $p = .07$ . This difference may be a consequence of output interference effects: the benefit of short retention interval may have been reduced to *B* items because they were only tested in the second half of the test list.



**Figure 4.2. ROC curves for A and B items (Exp. 5a).**

*A* items: ROC curves for *long* and *strong* lists lie slightly below the curves for *short* lists but not significantly so (no list-length and list-strength effects). *B* items: ROC curves for *strong* lists lie above the other curves showing that the strength manipulation was effective. There is no difference between *short* and *long* lists (no list-length effect). Data collapsed across retention intervals.  $N = 48$ .

Figure 4.2 shows the ROC curves for both *A* and *B* items. For *A* items, the ROC curves largely overlap (although the curves for *long* and *strong* lists are somewhat lower than the curve for *short* lists). This illustrates the null LLE and LSE in this experiment. For *B* items, the ROC curve for *strong* lists lies above the other curves, illustrating the increase in discriminability with item repetition. The curve for *long* lists lies somewhat above the curve for *short* lists, hinting at a negative LLE.

### Bias: *A* and *B* items

For *A* items, a 2 (retention interval: short, long)  $\times$  2 (list type: *short*, *long*, *strong*) within-participants ANOVA on the bias measure ( $c_a$ ) revealed no main effects and no interaction (all  $F$ s  $< 1.97$ , all  $p$ s  $> .15$ ). For *B* items, on the other hand, there was a strong shift in response bias across retention intervals as revealed by a 2 (retention interval: short, long)  $\times$  2 (list type: *short*, *long*, *strong*) ANOVA. Participants were more liberal in outputting “old” responses to *strong* lists than to *short* and *long* lists,  $F(2,90) = 8.86$ ,  $MSE = 0.01$ ,  $p < .001$ . The latter result seems to contradict data from false alarms, which showed no change across list types (and could thus be interpreted as evidence *against* a bias shift). The two results can be reconciled, however, by recalling that bias and false alarms measure different, albeit related, quantities and therefore may produce different results. The bias measure takes into account both hits and false alarms. Thus, one way of interpreting the results is by positing that the increase in bias was mostly caused by the significant increase in hits, due to the repetition of *B* items, with little or no change in false alarms. There was no difference in bias across items types (*A* and *B*) for *short* and *long* lists (all  $F$ s  $< 1.8$ , all  $p$ s  $> .19$ ). Table 4.3 presents these results.

**Table 4.3. Bias across item types ( $c_a$ ) (Exp. 5a).**

List type	A items				B items			
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>	
<b>Short</b>	-.06	$\tau$	$\tau$	0.04	-.01	$\tau$	$\tau$	0.05
<b>Long</b>	-.01	$\tau$	$\perp$	0.03	-.04	$\tau$	$\perp$	0.05
<b>Strong</b>	0.02	$\perp$	$\perp$	0.04	-.19	$\perp$	$\perp$	0.04

Note.  $c_a$  = bias (hits and false alarms at  $X_3$ ). A/B items = items from beginning of study list (in *strong* lists, *B* items are repeated). *n* non-significant; \*\*\*  $p \leq .001$ . Data collapsed across retention intervals (short and long).

### 4.2.3. Discussion

Experiment 5a yielded no LLE and no LSE. Both raw measures (hits and false alarms) and derived measures ( $A_z$ ) showed no reliable differences in discrimination of *A* items across *short*, *long* and *strong* lists. The null LSE, in particular, occurred despite the successful strengthening of *B* items in *strong* lists and despite the

presence of the ubiquitous between-list strength-based mirror effect. Moreover, there was no modulatory effect of retention interval on sensitivity across list types.

The result is consistent with the continuous memory assumption advanced by Murdock and Kahana (1993), according to which interference from prior study lists can shrink the LSE for recognition. The continuous memory assumption, however, still leaves room for a positive LLE, insofar as the effect can be generated by other factors (e.g., the forgetting parameter  $\alpha$  in TODAM). The fact that neither effect was observed here speaks against that possibility. In the next experiment, we try to reduce the effect of previous studied material by having participants studying 3 rather than 6 lists per experimental session.

### 4.3. Experiment 5b: Retention interval, without new, 3x, two

The null LLE and LSE in Experiment 5a suggest that having all lists studied and tested in one session may obscure any potential changes in sensitivity. Global matching models, like SAM and MINERVA2, predict LLE and LSE because the matching variance increases with list length and list strength, while the difference between the means of *target* and *lure* distributions remains the same. If extra-list variability is added to the matching process, however, changes in variability caused by length and strength manipulations may be masked.

To see how extra-list variability may mask changes in variance, let us assume that the ratio of *lure-to-target* variability in *short* lists is 0.8 ( $\sigma_D = 1$ ;  $\sigma_T = 1.25$ ). If length and strength manipulations increase the variance of *target* items in *long* and *strong* lists to, say,  $\sigma_T = 1.5$ , then the *lure-to-target* ratio would drop to 0.67, a 16% decrease in slope ratio across list types. Now consider the case when extra-list variance plays a role in the matching process by adding a fixed amount of variability (say 2) to both *lure* and *target* distributions. When that is the case, the slope ratio for *short* lists would rise to 0.92 ( $\sigma_D = 1 + 2$ ;  $\sigma_T = 1.25 + 2$ ) and the *lure-to-target* ratio in *long* and *strong* lists would fall to 0.86 ( $\sigma_D = 1 + 2$ ;  $\sigma_T = 1.5 + 2$ ), a 7% drop in slope ratio. Thus, extra-list variability reduced the differences in slope ratio across list types, despite the fact that *target* variability increased by the same amount in

both cases (i.e., from 1.25 to 1.5). To the extent that extra-list variability plays a role in the matching process, it will work against finding differences between list types.

In an attempt to reduce extra-list variability, thereby increasing the chances of observing an LLE and an LSE, we reran Experiment 5a with the only difference that retention interval was varied between *two sessions*, carried out on different days. Participants would thus be exposed to 3 lists per session rather than 6 in the current experiment. By having the sessions separated by at least one day, we also hoped to further reduce inter-list variability due to recent findings suggesting that sleep may help reduce interference between lists (Ellenbogen et al., 2006).

#### 4.3.1. Methods

##### Participants

Twenty-four University of Warwick students (12 males; age:  $M = 25.2$ ,  $SD = 6.8$ ) were tested individually and paid £6 to take part in the study.

##### Materials, Design and Procedure

The only difference between Experiments 5a and 5b is that in the latter participants attended two experimental sessions on different days rather than one session. In each session, participants completed four blocks (one practice, three experimental) in one of the two retention interval levels (either short or long interval). Order was counterbalanced across participants.

#### 4.3.2. Results

##### Hits and false alarms: A items

A 2 (word type: *target*, *SP lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) within-participants ANOVA on proportion of “old” responses revealed a strong main effect of word type,  $F(1,23) = 140.06$ ,  $MSE = 0.08$ ,  $p < .001$ , such that the proportion of “old” responses was higher for *targets* ( $M = .74$ ,  $SEM = 0.02$ ) than for *SP lures* items ( $M = .33$ ,  $SEM = 0.02$ ), confirming that discrimination was above chance. There was also an interaction between word type and list type,  $F(2,46) = 3.75$ ,  $MSE = 0.01$ , suggesting that hits decreased from *short* to *long* and from *short* to *strong* lists and false alarms increased in the opposite direction. All other main effects and interactions were not significant (all  $ps > .28$ ).

Separate two-way repeated-measures ANOVAs were carried out on the proportion of “old” responses for each word type (*target* and *SP lure*) with retention interval (*short* and *long*) and list type (*short*, *long* and *strong*) as the independent variables. For *targets*, there was an effect of list type,  $F(2,46) = 4.79$ ,  $MSE = 0.02$ , such that the hit rates were lower for *strong* and *long* lists. There was no main effect of retention interval and no interaction ( $ps > .49$ ). For *SP lures*, there were no significant main effects or interactions ( $ps > .75$ ). Results are presented in Table 4.4. Sensitivity ( $d'$ ) and bias ( $c$ ) for each retention interval are reported in Appendix 1. The results suggest that the interaction between word type and list type identified by the three-way ANOVA above was driven mainly by the decrease in hits across lists.

**Table 4.4. Hits and false alarms across item types (Exp. 5b).**

<b>A items</b>								
<b>List type</b>	<b>HR Targets</b>				<b>FAR SP lures</b>			
	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>
<b>Short</b>	.78	τ	τ	.02	.32	τ	τ	.03
		**	**			n	n	
<b>Long</b>	.71	τ	⊥	.03	.33	τ	⊥	.02
		n				n		
<b>Strong</b>	.71	⊥	⊥	.03	.34	⊥	⊥	.03
<b>B items</b>								
<b>List type</b>	<b>HR Targets</b>				<b>FAR SP lures</b>			
	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>
<b>Short</b>	.73	τ	τ	.03	.33	τ	τ	.03
		n				n	†	
<b>Long</b>	.70	τ	⊥	.02	.36	τ	⊥	.02
		***				*		
<b>Strong</b>	.85	⊥	⊥	.02	.27	⊥	⊥	.03

*Note.* HR = hits; FAR = false alarms; SP = switched plurality; A/B items = items from the beginning of the study list (in *strong* lists, *B* items are repeated). *n* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ . Data collapsed across retention intervals.  $N = 24$ .

There was an effect of list length, as the interaction between word type (*target* vs. *SP lure*) and list type (*short* vs. *long*) was significant,  $F(1,23) = 7.27$ ,  $MSE = 0.01$ . There was also an effect of list strength, as the interaction between word type and list type (*short* vs. *strong*) was significant,  $F(1,23) = 7.57$ ,  $MSE = 0.01$ . Thus, at the level of hits and false alarms, both length and strength manipulations were effective.

### Hits and false alarms: *B* items

A 2 (word type: *target*, *SP lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) within-participants ANOVA on proportion of “old” responses yielded a main effect of word type (more “old” responses to *targets*) and an interaction between list type and word type (more “old” responses to *targets* and fewer “old” responses to *SP lures* in *strong* lists than in *short* and *long* lists;  $F_s > 27.21$ ,  $p_s < .001$ ). The results are a consequence of the repetition of *B* items in *strong* lists and indicate a between-list strength-based mirror: the proportion of “old” responses in *strong* lists increased for *targets* and slightly decreased for *SP lures* compared to the proportion of “old” responses in *short* and *long* lists.

To confirm these results, two-way repeated-measures ANOVAs on the proportion of “old” entries were carried out for each word type (*target* and *SP lure*) with retention interval (*short* and *long*) and list type (*short*, *long* and *strong*) as the independent variables. For *targets*, there was a strong main effect of list type,  $F(2,46) = 21.18$ ,  $MSE = 0.20$ ,  $p < .001$ , such that there were more hits in *strong* lists, reflecting better memory for repeated *B* items. There was no effect of retention interval and no interaction ( $p_s > .82$ ). For *SP lures*, there was a main effect of list type,  $F(2,46) = 4.08$ ,  $MSE = 0.03$ ; post-hoc LSD analyses revealed that false alarms were lower to *strong* lists than to both *short* and *long* lists. The latter results confirm that the between-list strength-based mirror effect was indeed complete: hits in *strong* lists increased and false alarms decreased relative to the other list types. The result contrasts with the incomplete strength-based mirror effect in Experiment 5a. Hits and false alarms, collapsed across retention intervals, are shown in Table 4.4.

We also assessed whether there was an effect of length with *B* items. Although hits and false alarms behaved in a way consistent with an effect (hits decreased, false alarms increased), the changes were not large enough: the interaction between word type (*target* vs. *SP lure*) and list type (*short* vs. *long*) was not significant,  $F(1,23) = 2.19$ ,  $MSE = 0.02$ ,  $p = .15$ . Thus, the effect of length observed with *A* items was not replicated with *B* items. It is possible that the effect for *B* items might have been masked because *B* items were tested later (second half) in the test list.



### Hits and false alarms: *A* and *B* (within-list strength-based effects)

To assess whether the strength manipulation yielded a within-list strength-based mirror effect, we conducted a 2 (word type: *target*, *SP lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  2 (item strength: *A*, *B*) ANOVA on proportion of “old” only for *strong* lists. There was an interaction between word type and item strength such that proportion “old” for *targets* increased and for *SP lures* decreased from *A* (weak) to *B* (strong) items,  $F(1,23) = 26.02$ ,  $MSE = 0.02$ ,  $p < .001$ .

To confirm whether a mirror effect did in fact occur, two two-way ANOVAs on proportion “old” were conducted, one for each word type (*targets* and *SP lures*), with item strength and retention interval as the independent variables. The ANOVAs revealed that hits increased from *A* (weak) to *B* (strong) items ( $M_A = .71$ ,  $SEM = .03$ ;  $M_B = .85$ ,  $SEM = .02$ ;  $p < .001$ ) and that false alarms marginally decreased ( $M_A = .34$ ,  $SEM = .03$ ;  $M_B = .27$ ,  $SEM = .03$ ;  $p = .08$ ). The pattern suggests that participants may change their criterion within the same list when there is an abrupt and permanent change in test-item properties (e.g., weak to strong). The mirror pattern, however, was dominated by the increase in hits with item strength. Alternatively, the criterion may have remained unchanged, and the fall in SP false alarms may be accounted for by an increase in recall, which was used to reject lures.

### Sensitivity: *A* and *B* items

Sensitivity ( $A_z$ ) was estimated by fitting Gaussian models to confidence data. Only 2 of the 144 models [24 participants  $\times$  2 retention intervals (*short*, *long*)  $\times$  3 list types (*short*, *long* and *strong*)] for *A* items were rejected at the .05 level. The results below refer to the data of the remaining 22 participants.

A 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) ANOVA conducted on  $A_z$  revealed a main effect of list type,  $F(2,42) = 5.39$ ,  $MSE = 0.01$ ,  $p = .008$ . Post-hoc LSD tests showed that participants’ discrimination was worse in *strong* ( $M = .73$ ,  $SEM = .03$ ) than in *short* lists ( $M = .80$ ,  $SEM = .02$ ;  $p = .001$ ), worse in *long* ( $M = .75$ ,  $SEM = .03$ ) than in *short* lists ( $p = .04$ ) and approximately the same in *strong* and *long* lists ( $p = .37$ ). There was no effect of retention interval and no interaction ( $F_s < 1$ ,  $p_s > .40$ ). The effect of list type on  $A_z$  confirms the

presence of LLE and LSE in this experiment. Table 4.5 presents the results collapsed across retention intervals.

For *B* items, there were 2 poor fits to the 144 SDT models at the .05 level. Results refer to the data from the remaining 22 participants. A 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) ANOVA conducted on  $A_z$  revealed a main effect of list type,  $F(2,42) = 12.10$ ,  $MSE = 0.01$ ,  $p < .001$ , showing that discriminability for *B* items was better for *strong* than for *short* and *long* lists (due to the repetition of *B* items in *strong* lists). The effect indicates that the strength manipulation was effective. There was no effect of retention interval and no interaction ( $F_s < 1$ ,  $p_s > .77$ ).

**Table 4.5. Sensitivity ( $A_z$ ; SSP comparison) across item types (Exp. 5b).**

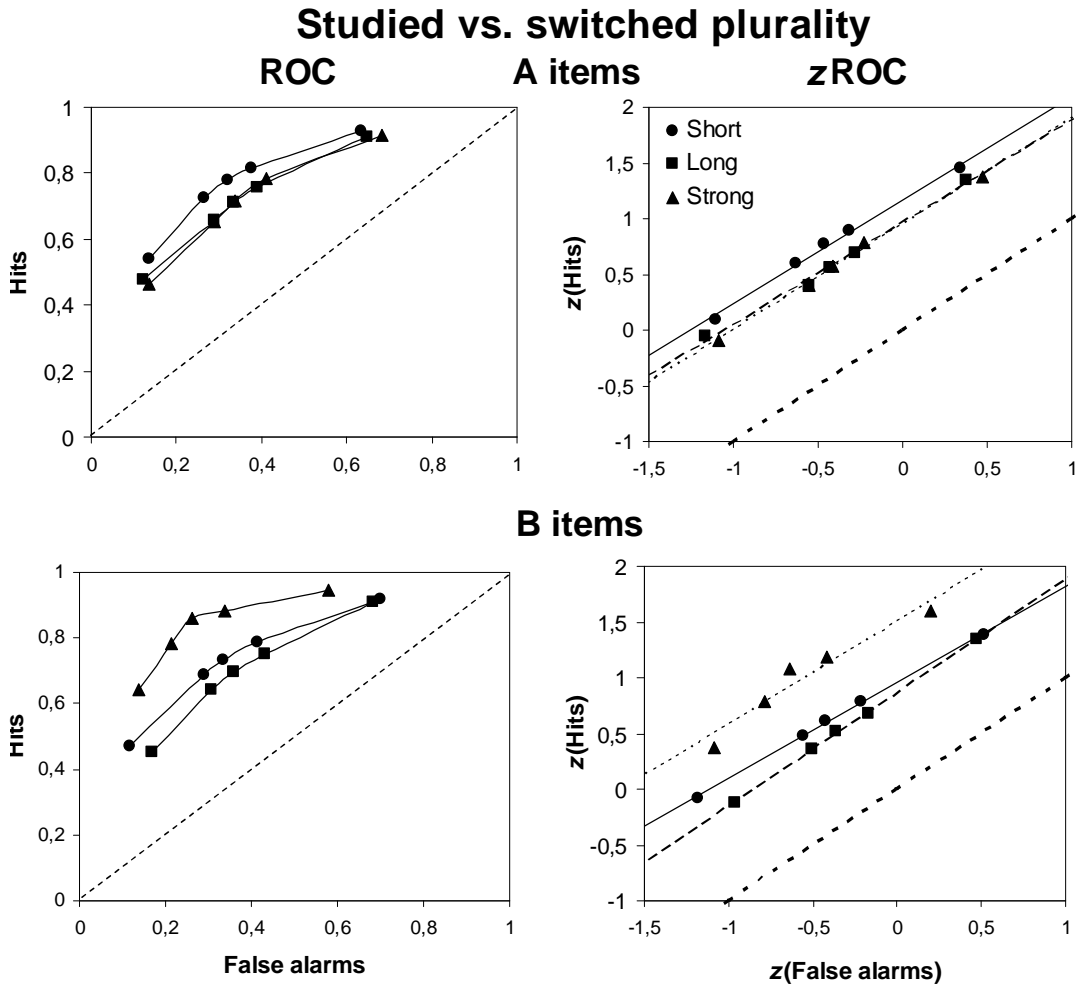
List type	A items				B items			
	<i>M</i>	<i>(N = 22)</i>		<i>SEM</i>	<i>M</i>	<i>(N = 22)</i>		<i>SEM</i>
<b>Short</b>	.80	τ	τ	.02	.77	τ	τ	.02
		*				†		
<b>Long</b>	.75	τ	⊥	***	.72	τ	⊥	***
		n				***		
<b>Strong</b>	.73	⊥	⊥	.02	.83	⊥	⊥	.03

*Note.*  $A_z$  = area under the ROC; A/B items = items from beginning of study list (in *strong* lists, *B* items are repeated). *n* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*\*  $p \leq .001$ . Data collapsed across retention intervals (short and long).

We further assessed whether the LLE found with *A* items would also be found with *B* items. This analysis is relevant because *B* items in *long* lists were treated at study exactly as *A* items and hence should also suffer list-length interference. A 2 (retention interval: *short*, *long*)  $\times$  2 (list type: *short*, *long*) showed a marginal effect of list type,  $F(1,21) = 3.24$ ,  $MSE = 0.01$ ,  $p = .09$ , replicating with *B* items the LLE observed with *A* items (although the effect size here was somewhat reduced).

When item type (*A* and *B*) was entered into a three-way ANOVA, the LLE result was further corroborated, as discrimination was significantly lower in *long* lists,  $F(1,20) = 4.65$ ,  $MSE = 0.02$ . In addition, there was a main effect of item type,  $F(1,20) = 6.07$ ,  $MSE = 0.01$ , whereby *A* items (tested in the first half of the test list;  $M_A = .78$ ,  $SEM = 0.02$ ) were better recognised than *B* items (tested in the second half of the test list;  $M_B = .75$ ,  $SEM = 0.02$ ). The result illustrates *output interference*: items tested later on a list are worse recognised than items tested earlier, regardless

of their original position at study. The result also lends some credence to the idea that the null LLE with hits and false alarms observed earlier could have been influenced by output interference effects.



**Figure 4.3. ROC and zROC curves for A and B items (Exp. 5b).**

*A items:* ROC and zROC curves for *long* and *strong* lists lie below the curves for *short* lists, showing list-length and list-strength effects. zROC slopes (all  $< 1$ ) did not differ across list types. *B items:* ROC and zROC curves for *strong* lists lie above the curves for *short* and *long* lists, showing that the strength manipulation was effective. ROC and zROC curves for *long* lists lie below the curves for *short* lists (list-length effect). zROC slopes did not differ significantly across list types, though there was a trend for higher slopes in *long* lists ( $\approx 1$ ). Data collapsed across retention intervals.  $N = 24$ .

Figure 4.3 depicts ROC and zROC curves for A and B items collapsed across retention intervals. For A items, ROC and zROC curves for *long* and *strong* lists lie below the curves for *short* lists, clearly illustrating an LLE and an LSE. Moreover, the zROCs were well fit by straight lines (all  $R^2$ s  $> .99$ ), consistent with the SDT assumption of underlying Gaussian distributions of familiarity. Finally, all zROC slopes were less than 1, consistent with the assumption that the *target* distribution

has greater variance than the *SP lure* distribution (unequal-variance SDT model). For *B* items, the ROC and  $z$ ROC curves for *strong* lists lie above the curves for *short* and *long* lists, showing that the strength manipulation was successful in increasing *B*-item sensitivity. Furthermore, ROC and  $z$ ROC curves for *long* lists lie below the curves for *short* lists, replicating with *B* items the LLE found with *A* items. Finally,  $z$ ROC slopes for *short* and *strong* lists were less than 1, suggesting that *target* and *SP-lure* variances were different, whereas the  $z$ ROC slope for *long* lists was approximately 1, suggesting that *target* and *SP-lure* variances were about the same (all  $R^2$ 's  $> .99$ , except for the  $z$ ROC of *strong*, *B* items, where  $R^2 = .94$ ).

### Bias: *A* and *B* items

For *A* items, a 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) within-participants ANOVA on the bias measure ( $c_a$ ) revealed no main effects and no interaction (all  $F$ s  $< 1.40$ , all  $p$ s  $> .26$ ). Similarly, there was no main effect of list type for *B* items as revealed by a 2 (retention interval: *short*, *long*)  $\times$  2 (list type: *short*, *long*, *strong*) ANOVA (all  $F$ s  $< 1.18$ , all  $p$ s  $> .32$ ), although there was a trend towards more liberal responses in *strong* lists ( $M_{bias} = -0.16$ ,  $SEM = 0.04$ ) than in *long* lists ( $M_{bias} = -0.05$ ,  $SEM = 0.05$ ;  $p = .08$ ). Table 4.6 presents these results.

**Table 4.6. Bias across item types ( $c_a$ ) (Exp. 5b).**

List type	A items				B items			
	$M$		$SEM$		$M$		$SEM$	
<b>Short</b>	-.17	$\tau$	$\tau$	0.06	-.08	$\tau$	$\tau$	0.06
<b>Long</b>	-.08	$\tau$	$\perp$	$n$	0.04	-.05	$\tau$	$\perp$
<b>Strong</b>	-.09	$n$	$\perp$	0.07	-.16	$\perp$	$\perp$	0.04

Note.  $c_a$  = bias (hits and false alarms at  $X_3$ ). A/B items = items from beginning of study list.  $n$  non-significant;  $\dagger p < .10$ . Data collapsed across retention intervals (short, long).

The non-significant shift in bias seems to contradict data from false alarms, which showed a change across list types (and hence could be construed as evidence for a bias shift). As with the previous experiment (5a), the discrepant results can be reconciled by noting that bias and false alarms may sometimes produce different results, since the bias index takes also hits into account. Thus, one can interpret the present results by arguing that bias remained unchanged here because the decrease

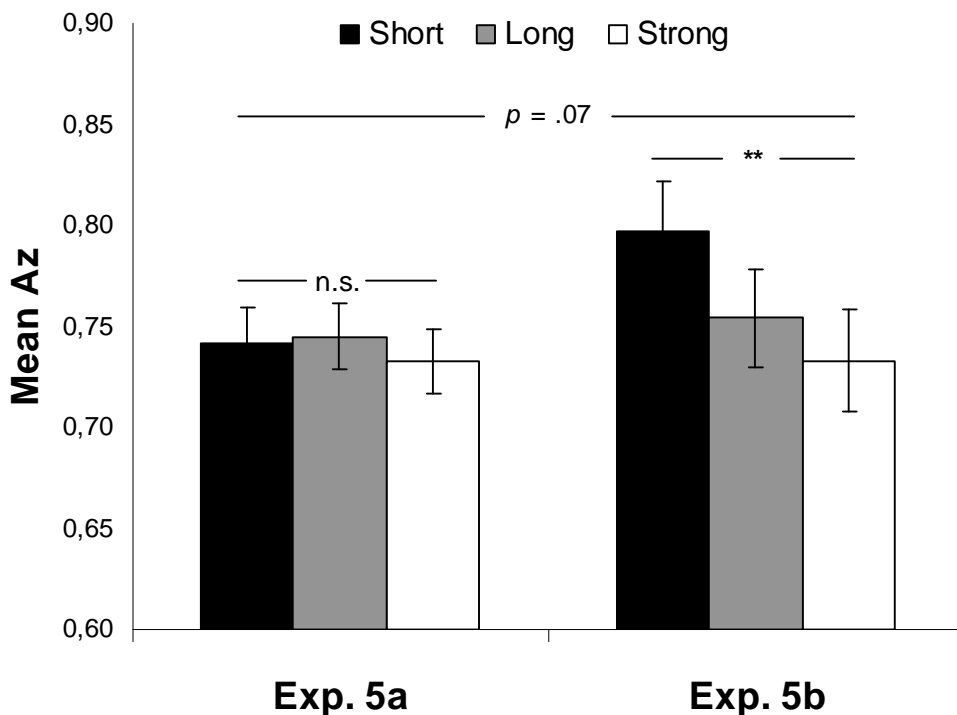
in false alarms (i.e., fewer “old” responses) for *B* items in *strong* lists was offset by an equivalent increase in hits (i.e., more “old” responses). In fact, the percent drop in false alarms from *short* to *strong* lists (18%) was similar to the percent rise in hits across lists (16%). There was no difference in bias across item types ( $M_A = -.09$ ,  $SEM = 0.04$ ;  $M_B = -.10$ ,  $SEM = 0.03$ ;  $p = .72$ ) and no interactions with retention interval and list types (all  $F_s \leq 1$ , all  $p_s > .36$ ).

### Experiment 5a vs. Experiment 5b (number of sessions)

In Experiment 5a, participants completed 6 study-test blocks within the same experimental session. In Experiment 5b, participants completed 6 study-test blocks in two sessions at least one day apart (3 blocks per session). In the former, no LLE and no LSE were found, whereas in the latter both an LLE and an LSE were found. In order to confirm this pattern statistically, it is important to evaluate whether the interaction between list type (*short*, *long*, *strong*) and number of sessions (1 in Exp. 5a vs. 2 in Exp. 5b) is significant. Another result that sets the experiments apart is the fact that retention interval affected overall sensitivity in Experiment 5a ( $A_z$  was higher when retention interval was shorter) but not in Experiment 5b. The only difference between the experiments, apart from the number of sessions, is the fact that there were more participants in the former ( $N = 48$ ) than in the latter ( $N = 24$ ).

A 2 [experiment: 5a (1 session), 5b (2 sessions)]  $\times$  3 (list type: *short*, *long*, *strong*)  $\times$  2 (retention interval: *short*, *long*) mixed-design ANOVA on sensitivity  $A_z$  revealed a marginal interaction between experiment and list type,  $F(2,136) = 2.76$ ,  $MSE = 0.01$ ,  $p = .07$ , weakly suggesting a differential role of number of experimental sessions across list types. There was a significant interaction between experiment and retention interval,  $F(1,68) = 4.63$ ,  $MSE = 0.01$ , confirming the differential effect of retention interval between experiments. Finally, there was an effect of list type across experiments,  $F(2,136) = 4.27$ ,  $MSE = 0.01$ . Post-hoc LSD tests revealed a marginally significant LLE (sensitivity was higher for *short* lists than for *long* lists,  $p = .09$ ) and a significant LSE (sensitivity was higher for *short* lists than for *strong* lists,  $p = .003$ ). The remaining main effects and interaction were not significant ( $F_s < 1$ ,  $p_s > .45$ ). Figure 4.4 illustrates the results with data collapsed across intervals.

We also tested whether the number of sessions had an effect on *target* variance by comparing, across experiments, the standard deviations of *target* distributions (relative to the standard deviation of *SP lures*, which was fixed at 1 in our models; see 2.3.2). A 2 [experiment: 5a (1 session), 5b (2 sessions)]  $\times$  3 (list type: *short*, *long*, *strong*)  $\times$  2 (retention interval: *short*, *long*) mixed-design ANOVA on estimated standard deviations revealed a marginal main effect of number of sessions,  $F(1,68) = 3.09$ ,  $MSE = 0.53$ ,  $p = .08$ , such that the standard deviation of *targets* (relative to *SP lures*) was higher when participants attended two sessions (Exp. 5b:  $M = 1.31$ ,  $SEM = 0.06$ ) than when they attended one session (Exp. 5a:  $M = 1.18$ ,  $SEM = 0.04$ ). The remaining terms were not significant ( $F_s < 1$ ,  $p_s > .59$ ). *Target* variance across list types and retention intervals was 1.25 times larger than the *lure* variance ( $M = 1.25$ ,  $SEM = 0.03$ ; i.e.,  $\sigma_D/\sigma_T = 1/1.25 = 0.8$ ), which is the value found in previous studies (Mickes et al., 2007; Ratcliff et al., 1992).



**Figure 4.4. Sensitivity across number of sessions (Exps. 5a / 5b).**

In Experiment 5a, where participants completed 6 study-test blocks in one session, no LLE and no LSE were found (i.e., no differences in sensitivity  $A_z$  between *short*, *long* and *strong* lists). In Experiment 5b, where participants completed 6 study-test blocks in two sessions (3 blocks per session on different days), both an LLE (drop in  $A_z$  for *long* lists relative to *short* lists) and an LSE (drop in  $A_z$  for *strong* lists relative to *short* lists) were found. Data collapsed across retention intervals; A items only. Error bars = SEM. *n.s.* non-significant; \*\*  $p < .01$ ;  $p$ -value on top bar represents interaction term (number of sessions  $\times$  list type).  $N = 70$ .

The result is consistent with the idea that higher inter-list variability, which is presumably larger during long experimental sessions, may have masked potential length and strength effects in Experiment 5a. The slope ratio results, however, should be interpreted with caution. First, there was no difference in standard deviations across list types, suggesting that increases in variance were not the driving force behind the observed sensitivity differences across list types. In fact, the bulk of the differences across lists can be accounted for by changes in the estimated mean familiarity values of the *target* distribution, as revealed by a two-way (retention interval  $\times$  list type) ANOVA [ $F(2,136) = 5.06$ ,  $MSE = 0.40$ ,  $p < .01$ ;  $M_{short} = 1.36 > M_{long} = 1.23 > M_{strong} = 1.11$ ;  $SEMs = 0.09$ ]. The result is in contrast with the prediction of some global matching models, according to which LLE and LSE should result from differences in target variance rather than target familiarity.

Another reason to be cautious about the difference in slope ratios reported here is that the estimates were obtained from too few trials per stimulus class (15 *targets*, 15 *SP lures*). MacMillan, Rotello and Miller (2004) showed that slope estimates were extremely variable and recommended samples of 200 trials per stimulus class in order to achieve acceptable levels of *accuracy* (i.e., difference between expected parameter estimate and true parameter value) and *precision* (i.e., degree of variability of parameter estimate). Finally, the difference in slope ratios may have been unduly influenced by extreme values: when the data is log-transformed to reduce the influence of extreme estimates, the difference in slopes goes away.

In sum, the results suggest that a long experimental session may reduce the likelihood of observing list-length and list-strength effects. LLE and LSE were found when participants underwent two sessions consisting of 3 study-test blocks each but not when they attended a single, hour-long session of 6 study-test blocks. Although the result is clear enough, the reason behind it is not. The hypothesis that extra-list variance plays a larger role during long experimental sessions, masking any potential changes in variance across list types, was not borne out by the data.

### 4.3.3. Discussion

Experiment 5*b* yielded reliable LLE and LSE with both raw measures (hits and false alarms) and derived measures ( $A_z$ ). That is, discrimination was lower for *long* and *strong* lists compared to *short* lists. This contrasts with results from Experiment 5*a*, where no differences in raw and derived measures were observed across list types. There was also no reliable effect of retention interval in both experiments.

It is important to note that the LSE has been observed without a concurrent shift in response bias. Hits fell and false alarms rose in *strong* lists relative to *short* lists. The result is relevant because most LSEs found so far have been accompanied by an upward shift in bias: participants become less likely to respond “old” to both *targets* and *lures* after studying *strong* lists. Van Zandt (2000) showed that changes in bias  $c$  may affect sensitivity  $d'$ . She manipulated the proportion of *targets* present at test and found that, as participants became more conservative (i.e., fewer “old” responses),  $d'$  decreased. Thus, it could be argued that a list-strength effect is just a by-product of criterion shifts:  $d'$  falls in *strong* lists not because discriminability is impaired;  $d'$  falls because participants are more conservative when responding to *strong* lists. The fact that  $d'$  significantly decreased in this experiment without a concurrent change in  $c$  (see Appendix 1 for  $d'$  measures of LSE) provides strong evidence against the criterion-shift view. Although we have been using  $A_z$  as a sensitivity measure in order to avoid Van Zandt’s (2000) criticism, it is reassuring that an LSE has been found here with a single-point sensitivity measure such as  $d'$ .

The size of the interference effects in Experiment 5*b* (LLE:  $d_z = 0.40$ ; LSE:  $d_z = 0.74$ ) were greater than in Experiment 3 (SSP comparison; LLE:  $d_z = 0.09$ ; LSE:  $d_z = 0.42$ ). The LSE in Experiment 5*b* was also greater than the LSE in Norman (2002, Exp. 2; SSP comparison:  $d_z = 0.44$ ), where the strong items were presented 6 times. The LLE in the current experiment was similar in size to the LLE in Cary and Reder (2003, Exp. 3;  $d'$  measure:  $g = 0.46$ ), where the *long* list was four times longer than the *short* list. The results indicate that forcing participants to rely on recollection to make their recognition decisions (i.e., having only *targets* and *SP lures* at test) may in fact further increase the magnitude of list-length and list-strength effects.



The shapes of ROC and  $z$ ROC curves in Figure 4.3 contain potentially informative features. The fact that the ROC curves were concave and the  $z$ ROC curves were linear suggests that the underlying strength-of-evidence distribution is Gaussian. The result appears to contradict the predictions of some dual-process models (e.g., Yonelinas, 2001), according to which recollection is assumed to be *all-or-none*, thereby being best described by a threshold function. In such models, recollection ROCs are predicted to be linear and  $z$ ROCs are predicted to be U-shaped. These predictions, however, were not supported by our data. In fact, when we try to fit the pooled ROC data for *A* items in Experiment 5*b* with a high-threshold model, the fits are extremely poor. Whereas the worse unequal-variance SDT model across list types is able to fit the data at an acceptable level [ $\chi^2(4) = 5.46, p = .14$ ], the best high-threshold model was nowhere close to fitting the data appropriately [ $\chi^2(4) = 150.65, p = 10^{-31}$ ]. Thus, it seems that participants refrain from using an all-or-none strategy at test even when familiarity by itself is not diagnostic of an item being old.

By contrast, ROC and  $z$ ROC results suggest that a continuous recollection process may be operating at test. The curves in Experiment 5*b* replicate qualitatively the pattern reported by Heathcote et al. (2006, Exp. 4), who fitted their results with a model that assumed continuous recollection. Moreover, the curves are similar to the curves obtained in source memory (Slotnick, Klein, Dodson, & Shimamura, 2000) and associative recognition (Kelley & Wixted, 2001) studies, suggesting that a similar, continuous recollection process may underlie those different tasks. There are reasons, however, to be cautious about conclusions drawn from the shapes of ROC and  $z$ ROC curves. First, Lockhart and Murdock (1970) pointed out that many unimodal distributions produce linear curves in  $z$ ROC space. Thus, the fact that a  $z$ ROC is linear does not entail that the underlying distribution is Gaussian; it could be a Gamma distribution instead. Second, Malmberg (2002) argued that, although threshold models do predict linear ROC curves when the curves are constructed from old-new data (see 2.2.2), threshold models do not necessarily predict linear ROCs when the curves are constructed from confidence ratings. A threshold model can also generate concave ROCs depending on the mapping between thresholds (discrete states) and confidence ratings (responses). Thus, the fact that in this experiment the ROCs were concave does not conclusively rule out the possibility that the underlying distribution is discrete and that recollection is all-or-none.

Another feature worth noting about the ROCs in Figure 4.3 is that the distance between the 5 criteria is shorter for conditions where discrimination is higher. For *A* items, the criteria for *short* lists are closer together than the criteria for *long* and *strong* lists. For *B* items, the criteria for *strong* lists are closer than the criteria for *short* and *long* lists. The tendency for response criteria to come closer in conditions where discrimination is higher (or, alternatively, to fan out when discriminability is lower) has been previously reported by Stretch and Wixted (1998a) and suggests that participants attempt to maintain a similar odds ratio for a given confidence rating across conditions of varying difficulty. For example, if a test item in the high-discriminability condition elicits a familiarity signal that is 10 times more likely to have come from the *target* distribution than from the *SP lure* distribution, and if the participant endorses that signal with high-confidence (i.e., responds *definitely old*), then the same participant tends to adjust that confidence rating in order to keep roughly the same 10:1 odds ratio in a condition where discriminability is lower. To keep that odds ratio, criteria need to fan out when *target* and *lure* distributions come closer (see Stretch & Wixted, 1998a, Fig. 3, for a graphical illustration). Intuitively, participants require more evidence to output a *definitely old* response in a situation where discriminability is poor presumably because personal past experience showed that, in those situations, high-confidence mistakes (e.g., false alarms) may be costly. A similar reasoning applies to high-confidence *new* responses. Consequently, the criterion for *definitely old* responses shifts right on the familiarity axis and the criterion for *definitely new* responses shifts left when participants move from a high-discriminability to a low-discriminability situation.

There is one unexplained feature among the *z*ROC curves in Figure 4.3. The *z*ROC for *strong* lists (*B* items) appears to be curved downward rather than being straight. In fact, the *z*ROC is fitted significantly better when a quadratic component is added (goodness-of-fit  $\chi^2$  test of nested models;  $\chi^2(1) = 8.68, p = .003$ ). Such concave *z*ROC have been previously interpreted as noise in the data (Ratcliff et al., 1994, Fig. 11): when noise is added to 5% of the responses across all confidence ratings, a previously linear *z*ROC becomes concave. Concave *z*ROCs may also be produced by non-Gaussian distributions. In Yonelinas's (1994, 2001) dual-process model, the recollection component causes the *z*ROC to become U-shaped as it adds a large

amount of high-confidence old responses to the curve, pushing up its leftmost point. The interpretation of the concave  $z$ ROC in Figure 4.3, however, is complicated by the fact that it represents responses to strong items. Those items should elicit large amounts of recollection, which should cause the  $z$ ROC to bend upwards, not downwards. And if noise in the responses was responsible for the concave  $z$ ROC, then it is unclear why it was prominent only in *strong* lists; *short* and *long* lists were also presented to the same participants and yet their  $z$ ROCs were linear.

Neither sensitivity nor response bias were affected by retention interval. This was unexpected, given the significant modulatory effect of retention interval observed in Experiments 2 and 3. The null result is even more surprising given that models based on very different assumptions converge on the same prediction that shorter retention intervals should increase interference effects. CLS predicts that shorter retention intervals mean stronger connection weights for strong items and thus more interference on weak items through long-term depression on their discriminative features. Alternatively, BCDMEM predicts that short retention intervals discourage the reinstatement of the study context at test and, consequently, degrade performance to a larger extent than long retention intervals do.

It is possible that the time between beginning and end of the study list was not long enough to allow substantial changes in study context. Hindering contextual reinstatement at test, by means of a short retention interval, should have less of an impact on performance when context drifts little than when it drifts a lot. One way of increasing the likelihood of contextual drift is by increasing the size of the study list, thereby widening the gap between start and end of the study list. We sought to test this possibility, in the next experiment.

#### **4.4. Experiment 6: Retention interval, without new, 6x, two**

In the previous experiment, although the effect sizes of list-length and list-strength effects were higher than the effects obtained in the experiments in Chapter 1, there was no modulation of the effects by retention interval. It is possible that the design adopted in Chapter 4 requires stronger manipulations to elicit retention interval

effects. To increase the chances of observing an effect, participants in this experiment were tested in *two sessions* (on different days) to reduce inter-list interference. Moreover, the length and strength manipulation were incremented: the number of presentations of strong items rose from 3 to 6 and the long-to-short list-length ratio increased from 2:1 (120 vs. 60 items) to 3.5:1 (210 vs. 60 items). By varying retention interval and manipulation strength, we can also address some conflicting predictions made by current computational models.

According to BCDMEM (Dennis & Humphreys, 2001), the increase in list size should facilitate the detection of retention interval effects because a longer study-test lag provides more opportunities for the features present in the original study context to be lost due to temporal drift. To the extent that the context reinstated at test mismatches the original context present at study, performance should suffer. Study-test lag is 2.5 min. longer here than in Experiments 5*a* and 5*b*. Thus, BCDMEM predicts a drop in sensitivity across list types in this experiment. Crucially, this drop should be larger when retention interval is short because participants may decide to stick to the just-experienced context at the end of the list to anchor their responses, thereby reducing the ability to discriminate *targets*, which were studied at the beginning of the list, from *SP lures*, which were not studied. When retention interval is long, however, BCDMEM predicts no difference between *short*, *long* and *strong* lists: all three list types share the same study-test lag and context reinstatement should, at most, be only mildly impaired in *long* and *strong* lists relative to *short* lists. Moreover, studying extra items or repeating the same item should not affect the item-to-context mappings of other, previously studied items because all items, *target* and *interference*, are stored separately in the model.

By contrast, CLS (Norman & O'Reilly, 2003) predicts interference effects in both retention interval conditions. Interference should be higher at short intervals because the connection weights of *interference* items in the cortical and hippocampal models would not have decayed much with time, and high weights for *interference* items amount to less activation of discriminative features for the other, *target* items. For example, studying different types of “round fruits” will increase the activation in the model of shared features, such as “roundness”, but it will concurrently decrease the activation of specific features of a given target fruit (e.g., “round and red”). When

retention interval is long, however, the weights of *interference* items will have decayed more, thereby causing less disruption during the recognition test.

CLS also predicts an increase in both LLE and LSE with stronger manipulations. Adding even more items to a list or repeating the same items further should cause additional weakening of connections to discriminative features of *target* items. This contrasts with REM (Shiffrin & Steyvers, 1997), which predicts that LLE should increase with longer lists when *lures* are randomly similar to *targets* but not when *lures* are highly similar to *targets* and that LSE should remain unchanged (and negligible) with stronger lists regardless of lure similarity.

In REM, longer list lengths provide more opportunities for *lures* to match stored traces spuriously (causing a rise in false alarms) and more opportunities for targets to mismatch other items in memory (causing a fall in hits). Hence, hits and false alarms move in opposite directions, and the model predicts that LLE should rise with longer lists. When *lures* are very similar to *targets*, however, hits and false alarms move together: the fall in hits is shadowed by a fall in false alarms. Hence, no LLE is predicted in REM when *lures* are highly similar, regardless of list length.

Similarly, stronger lists should not alter the null LSE in REM. Repeating a study item causes its representation in REM to become more accurate (i.e., more features of the item are encoded in memory and the values of those features are more likely to be correctly stored). Because more features are stored, the match between a test item and its memory representation increases with repetition. Conversely, the match between a strong trace and a test item other than itself decreases. That is because the presence of more non-zero features in the strong trace provides more chances for a mismatch, contributing to a fall in the odds signal elicited by the test item. In other words, a strong trace is less confusable with a test item other than itself. In lists of mixed strength, strengthening some items (*interference* items) has the effect of decreasing the odds that a test item, both *target* (non-strengthened) and *lure* (not on the list), will surpass the response threshold because their match against the stored *interference* items will become lower and lower with their increasing differentiation. And since the odds decrease in tandem for *targets* and *lures*, recognition sensitivity

remains unaffected. Hence, REM predicts that no LSE should be observed even with an increase in the number of presentations of strong items.<sup>4</sup>

#### 4.4.2. Methods

##### Participants

Forty-eight University of Warwick undergraduates (16 males; mean age = 24.4, *SD* = 7.0) participated in the study. The experiment lasted 60 min. (two 30-min. sessions on different days) and participants were paid £6.

List type	Short retention interval		
	Study	Distractor	Test
Short	[AB]	257.5 s	[tA, spA] [tB, spB]
Long	[AB] [CDEFG]	10 s	[tA, spA] [tB, spB]
Strong	[AB] [BBBBB]	10 s	[tA, spA] [tB, spB]
	Long retention interval		
	Study	Distractor	Test
Short	[AB]	367.5 s	[tA, spA] [tB, spB]
Long	[AB] [CDEFG]	120 s	[tA, spA] [tB, spB]
Strong	[AB] [BBBBB]	120 s	[tA, spA] [tB, spB]

**Figure 4.5. Design of Experiment 6.**

A-G = matched groups of 30 words; [X,Y] = word groups X and Y are merged and word order in the resulting list is randomised; tA = targets from group A; spA = switched-plurality lures from group A; tB = targets from group B; spB = switched-plurality lures from group B.

<sup>4</sup> Amy Criss (personal communication, February, 2008) kindly ran some simulations of the REM model and confirmed that, under the parameter values used in the simulations, increasing the number of presentations of strong items (2x to 4x) makes no difference to the prediction of a null LSE.

## Materials

The word stimuli were made up of 354 nouns from the MRC Psycholinguistic Database (mean imageability = 5.73; concreteness = 5.79; familiarity = 5.11; frequency = 17.4 occurrences per million; length = 5.34). Nouns' plurals were generated by adding *s* to their singular form. Half the nouns were in singular form; half were in plural form. The words were assessed so that no items were strongly related to one another.

Twenty four words were used as fillers and the remaining 330 words were randomly assigned to 11 groups of 30 words, matched for word characteristics. Words were classified as *targets* (*A* or *B*; if presented at study and at test in the same plurality), *SP lures* (*A* or *B*; if presented at study and at test with their plurality reversed) or *interference* (if presented at study but not at test). *SP lures* were constructed from half of the *targets* by reversing their plurality. Of the 11 groups of 30 words, 6 groups consisted of *targets/SP lures* (3 groups of *A items*; 3 groups of *B items*) and 5 groups consisted of *interference* items (groups *C* to *G* in Figure 4.5). Different word samples were produced for each participant (one for each retention interval) and different samples were produced across participants (to balance the assignment of words to word groups and list types).

## Design and Procedure

Figure 4.5 illustrates the experimental design. Design and Procedure were identical to Experiment 5*b*, except that *long* lists were longer and *strong* lists were stronger. The list-length ratio increased from 2:1 in Experiment 5*b* to 3.5:1 in this experiment (*short* lists = 60 words; *long* lists = 210 words). The number of presentations of *B* items in *strong* list increased from 3 in Experiment 5*b* to 6 presentations in this experiment (*strong* lists = 60 different words, 210 study trials). Short retention interval was 257.5 s for *short* lists and 10 s for *long* and *strong* lists. Long retention interval was 367.5 s for *short* lists and 120 s for *long* and *strong* lists.

### 4.4.3. Results

### Hits and false alarms: A items

A three-way repeated-measures ANOVA on proportion of “old” responses with item type (*target* vs. *SP lure*), retention interval (*short* vs. *long*) and list type (*short*, *long* and *strong*) as independent factors yielded a main effect of word type,  $F(1,47) = 156.47$ ,  $MSE = 0.13$ ,  $p < .001$ , such that “old” responses were given more often to *targets* ( $M = .72$ ,  $SEM = 0.01$ ) than to *SP lures* ( $M = .35$ ,  $SEM = 0.02$ ). There was also a strong main effect of list type,  $F(2,94) = 11.25$ ,  $MSE = 0.02$ ,  $p < .001$ , showing that “old” responses were less frequent for *strong* lists than for both *short* and *long* lists. In addition, there was a main effect of retention interval,  $F(1,47) = 12.27$ ,  $MSE = 0.03$ ,  $p = .001$ , indicating that participants responded “old” less often when the interval was short. Finally, there was a marginal interaction between word type and list type,  $F(2,94) = 2.55$ ,  $MSE = 0.02$ ,  $p = .08$ , suggesting that hits and false alarms behaved differently depending on the list manipulation (hits were lower in *long* and *strong* lists; false alarms were lower in *strong* lists and higher in *long* lists). All other main effects and interactions did not reach significance ( $ps > .41$ ).

Separate two-way repeated-measures ANOVAs were carried out on the proportion of “old” responses for each word type (*target* and *SP lure*) with retention interval (*short* and *long*) and list type (*short*, *long* and *strong*) as the independent variables. For *targets*, there was a strong effect of list type,  $F(2,94) = 8.54$ ,  $MSE = 0.02$ ,  $p < .001$ , such that the hit rates were lower for *strong* lists. There was also a main effect of retention interval,  $F(1,47) = 9.59$ ,  $MSE = 0.02$ ,  $p = .003$ , showing that hits were lower when retention interval was short. There was, however, no significant interaction between list type and retention interval ( $p = .17$ ). For *SP lures*, there was a main effect of list type,  $F(2,94) = 5.83$ ,  $MSE = 0.02$ ,  $p = .004$ , showing that false alarms were higher for *long* lists and lower for *strong* lists compared to *short* lists. There was also a main effect of retention interval,  $F(1,47) = 5.61$ ,  $MSE = 0.02$ , such that there were fewer false alarms overall when retention interval was short. There was no interaction between list type and retention interval ( $p = .94$ ). The results suggest that the marginal interaction between word type and list type identified by the three-way ANOVA above was driven both by a decrease in hits for *long* and *strong* lists relative to *short* lists and by opposite effects of list type on false alarms (increase in *long* and decrease in *strong* lists). The results also indicate that



participants were more wary of responding “old” with short retention intervals, suggesting a shift in response criterion. Results are presented in Table 4.7. Sensitivity ( $d'$ ) and bias ( $c$ ) for each retention interval are reported in Appendix 1.

There was an effect of list length across retention intervals, as the interaction between word type (*target* vs. *SP lure*) and list type (*short* vs. *long*) was significant,  $F(1,47) = 4.88$ ,  $MSE = 0.02$ . There was no effect of list strength across retention intervals; the interaction between word type and list type (*short* vs. *strong*) was not significant,  $F(1,47) = 2.57$ ,  $MSE = 0.02$ ,  $p = .12$ . Thus, at the level of hits and false alarms, there was an effect of list length without a concurrent effect of list strength. Although the three-way interaction between word type, list type and retention interval was not significant, separate analyses at each retention interval revealed a trend suggesting that length and strength manipulations affected hits and false alarms more when retention interval was short. The  $p$ -values of the *word type*  $\times$  *list type* interaction terms for long and short retention intervals were, respectively, .23 vs. .03 for *long* lists, and .51 vs. .09 for *strong* lists.

#### Hits and false alarms: B items

A 2 (word type: *target*, *SP lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) within-participants ANOVA on proportion of “old” responses revealed a main effect of word type (more “old” responses to *targets*) and an interaction between list type and word type (more “old” responses to *targets* and fewer “old” responses to *SP lures* in *strong* lists than in *short* and *long* lists;  $F_s > 50.45$ ,  $p_s < .001$ ). These results, which follow from the repetition of *B* items in *strong* lists, show a between-list strength-based mirror: the proportion of “old” responses in *strong* lists increased for *targets* and decreased for *SP lures* compared to the proportion of “old” responses in *short* and *long* lists. There was also a main effect of list type,  $F(1,47) = 10.78$ ,  $MSE = 0.02$ ,  $p < .001$ , showing that participants responded “old” more often in *strong* lists than in both *short* and *long* lists. There was no main effect of retention interval and no interactions between retention interval and the other two variables ( $F_s < 1.27$ ,  $p_s > .28$ ).

To confirm the pattern above, two-way repeated-measures ANOVAs on the proportion of “old” responses were carried out for each word type (*target* and *SP*

*lure*) with retention interval (*short* and *long*) and list type (*short*, *long* and *strong*) as the independent variables. For *targets*, there was a strong main effect of list type,  $F(2,94) = 58.67$ ,  $MSE = 0.20$ ,  $p < .001$ , such that there were more hits in *strong* lists, showing that repeated *B* items were better recognised than non-repeated *B* items. There was no effect of retention interval and no interaction ( $ps > .52$ ). For *SP lures*, there was a main effect of list type,  $F(2,94) = 9.90$ ,  $MSE = 0.02$ ; post-hoc LSD analyses revealed that false alarms were lower to *strong* lists than to both *short* and *long* lists. The latter results confirm that the between-list strength-based mirror effect as hits in *strong* lists increased and false alarms decreased relative to the other list types. Table 4.7 shows hits and false alarms collapsed across retention intervals.

There was no effect of list length with *B* items across retention intervals as the interaction between word type (*target* vs. *SP lure*) and list type (*short* vs. *long*) was not significant,  $F(1,47) = 1.81$ ,  $MSE = 0.02$ ,  $p = .19$ . Although the false alarms did reliably increase from *short* to *long* lists (.34 vs .38, respectively), hit rates did not vary across lists (.73 for both list types). As in Experiment 5(a,b), effects of length could have been masked because *B* items were tested in the second half of the list.

**Table 4.7. Hits and false alarms across item types (Exp. 6).**

<b>A items</b>								
<b>List type</b>	<b>HR Targets</b>				<b>FAR SP lures</b>			
	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>
<b>Short</b>	.76	⌈	⌈	.02	.36	⌈	⌈	.03
		<i>n</i>				<i>n</i>	†	
<b>Long</b>	.73	⌈	⊥ ***	.02	.39	⌈	⊥	.02
		*				***		
<b>Strong</b>	.68	⊥	⊥	.02	.31	⊥	⊥	.03
<b>B items</b>								
<b>List type</b>	<b>HR Targets</b>				<b>FAR SP lures</b>			
	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>
<b>Short</b>	.73	⌈	⌈	.02	.34	⌈	⌈	.03
		<i>n</i>				*	*	
<b>Long</b>	.73	⌈	⊥ ***	.02	.38	⌈	⊥	.02
		***				***		
<b>Strong</b>	.91	⊥	⊥	.01	.29	⊥	⊥	.03

*Note.* HR = hits; FAR = false alarms; SP = switched plurality; A/B items = items from the beginning of the study list (in *strong* lists, *B* items are repeated). *n* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*\*  $p \leq .001$ . Data collapsed across retention intervals.  $N = 48$ .

### Hits and false alarms: *A* and *B* (within-list strength-based effects)

A 2 (word type: *target*, *SP lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  2 (item strength: *A*, *B*) ANOVA on proportion of “old” responses was conducted only for *strong* lists to assess whether the strength manipulation yielded a within-list strength-based mirror effect. The results showed strong main effects of item strength (more “old” responses to *B* items;  $p < .001$ ), word type (more “old” responses to *targets*;  $p < .001$ ) and retention interval (fewer “old” responses when the interval was short;  $p = .02$ ). There was an interaction between item strength and retention interval,  $F(1,47) = 6.32$ ,  $MSE = 0.02$ : the proportion of “old” responses decreased from long to short retention intervals only to *A* items. More importantly for the present purposes, there was an interaction between word type and item strength such that proportion “old” for *targets* increased and for *SP lures* decreased from *A* (weak) to *B* (strong) items,  $F(1,57) = 99.08$ ,  $MSE = 0.02$ ,  $p < .001$ .

To confirm whether the latter interaction indeed represents a mirror effect, two-way ANOVAs on proportion “old” were conducted for each word type (*targets* and *SP lures*), with item strength and retention interval as the independent variables. Hits increased from *A* (weak) to *B* (strong) items ( $M_A = .68$ ,  $SEM = .02$ ;  $M_B = .91$ ,  $SEM = .01$ ;  $p < .001$ ) but false alarms did not significantly decrease ( $M_A = .31$ ,  $SEM = .03$ ;  $M_B = .29$ ,  $SEM = .03$ ;  $p = .33$ ). The result does not replicate the marginally significant within-list decrease in false alarms observed in Experiment 5*b*, although there was a clear trend towards fewer *SP* false alarms for *B* items. As in Experiment 5*b*, the within-list mirror pattern observed here is dominated by an increase in hits with item strength but only a slight decrease in false alarms.

### Sensitivity: *A* and *B* items

Sensitivity  $A_z$  was estimated by fitting Gaussian models to confidence data. Of the 288 models fitted [48 participants  $\times$  2 retention intervals (*short* and *long*)  $\times$  3 list types (*short*, *long* and *strong*)], 2 were excluded due to poor fits ( $\chi^2 > .05$ ). Results refer to estimates of the remaining 46 participants whose data were fitted by the model across all three discrimination types. Table 4.8 summarises the results.

A 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) ANOVA conducted on  $A_z$  revealed a marginal effect of list type,  $F(2,90) = 2.97$ ,  $MSE = 0.01$ ,  $p = .06$ . Post-hoc LSD comparisons showed that sensitivity was worse in *strong* ( $M = .73$ ,  $SEM = 0.02$ ) than in *short* lists ( $M = .77$ ,  $SEM = 0.02$ ;  $p = .04$ ), worse in *long* ( $M = .73$ ,  $SEM = 0.02$ ) than in *short* lists ( $p = .04$ ) and did not differ between *strong* and *long* lists ( $p = .91$ ). The results confirm the presence of an LLE and an LSE in this experiment, replicating the effect found in Experiment 5*b*. There was no main effect of retention interval and no interactions ( $F_s < 1.2$ ,  $p_s > .29$ ).

For *B* items, there were 3 poor fits to the 288 SDT models. Results refer to data from the remaining 45 participants. A 2 (retention: *short*, *long*)  $\times$  3 (list: *short*, *long*, *strong*) ANOVA conducted on  $A_z$  revealed a main effect of list type,  $F(2,88) = 52.36$ ,  $MSE = 0.01$ ,  $p < .001$ , showing that sensitivity for *B* items was higher in *strong* than in *short* and *long* lists (due to the repetition of *B* items in *strong* lists). There was no effect of retention interval and no interaction ( $F_s < 2.29$ ,  $p_s > .11$ ).

We further evaluated whether the LLE found with *A* items was also present with *B* items. This would be expected as *A* and *B* items in *long* lists were treated at study in the same way and hence should both suffer interference with increasing list length. A 2 (retention: *short*, *long*)  $\times$  2 (list: *short*, *long*) showed that discrimination for *B* items in *long* lists was worse than in *short* lists,  $F(1,44) = 4.32$ ,  $MSE = 0.01$ , replicating the LLE observed with *A* items. There was also a marginal main effect of retention interval,  $F(1,44) = 2.88$ ,  $MSE = 0.01$ ,  $p = .09$ , weakly suggesting that overall discrimination across list types was worse when retention interval was short ( $M = .72$ ,  $SEM = .02$ ) compared to when it was long ( $M = .75$ ,  $SEM = .02$ ).

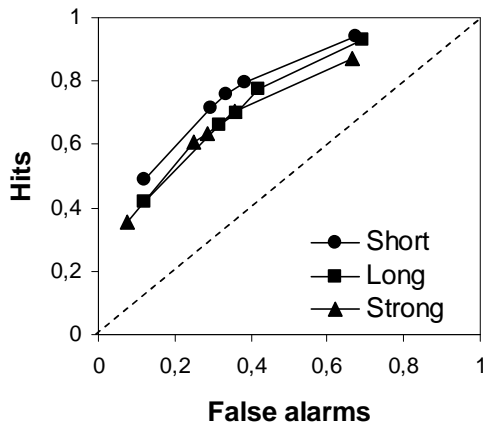
**Table 4.8. Sensitivity ( $A_z$ ; SSP comparison) across item types (Exp. 6).**

List type	A items			B items		
	<i>M</i>	<i>N</i> = 46		<i>M</i>	<i>N</i> = 45	
			<i>SEM</i>			<i>SEM</i>
<b>Short</b>	.76	⌈ *	.02	.76	⌈ *	.02
<b>Long</b>	.73	⌈ n	.02	.72	⌈ ***	.02
<b>Strong</b>	.73	⌊ ⌊	.02	.88	⌊ ⌊	.02

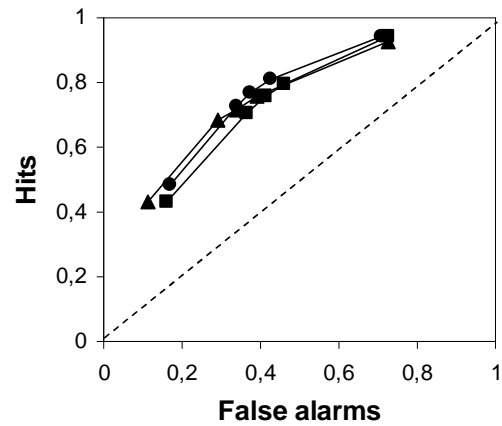
*Note.*  $A_z$  = area under the ROC; A items = targets; B items = interference items (presented 6x in *strong* lists). *n* non-significant; \*  $p < .05$ ; \*\*\*  $p < .001$ . Data collapsed across retention intervals.

## A items

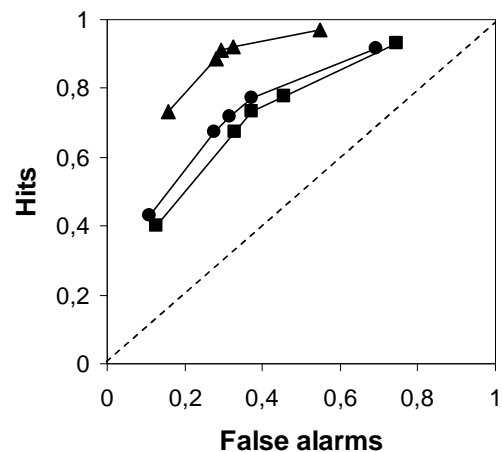
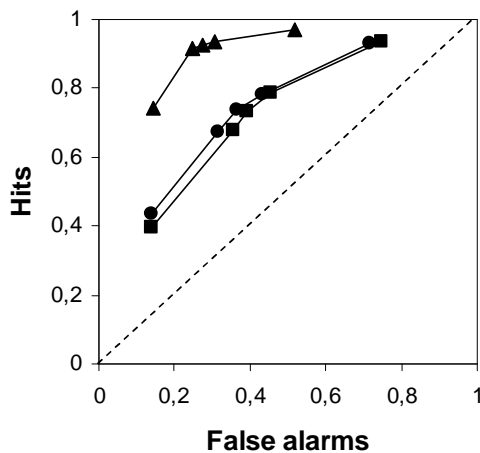
## Short retention interval



## Long retention interval



## B items



**Figure 4.6. ROC curves across retention intervals (Exp. 6).**

*A items*: ROC curves for *long* and *strong* lists lie below the curves for *short* lists (list-length and list-strength effects) when retention interval is short (10 s) but not when it is long (120 s). *B items*: ROC curves for *strong* lists lie above the other two curves, confirming that the strength manipulation was effective. The curves for *long* lists lie slightly below the curves for *short* lists, replicating the list-length effect observed for *A items*.  $N = 48$ .

Figure 4.6 depicts ROC curves for *A* and *B* items for each retention interval. For *A* items, the ROC curves for *long* and *strong* lists fall below the curves for *short* lists, illustrating the LLE and the LSE. This pattern is more evident when retention interval is short. When retention interval is long, the curves largely overlap. For *B* items, the ROC curves for *strong* lists fall above the curves for *short* and *long* lists, showing that the strength manipulation was successful with both retention intervals. In addition, ROC curves for *long* lists lie mostly below the curves for *short* lists in both retention intervals, replicating with *B* items the LLE found with *A* items.

### Bias: A and B items

For *A* items, a 2 (retention interval: short, long)  $\times$  3 (list type: *short*, *long*, *strong*) repeated-measures ANOVA on the bias measure ( $c_a$ ) revealed a strong main effect of list type,  $F(2,90) = 10.91$ ,  $MSE = 0.01$ ,  $p < .001$ , such that participants were more conservative in responding to *strong* lists than to both *short* and *long* lists. The ANOVA also revealed a marginal main effect of retention interval,  $F(1,45) = 3.69$ ,  $MSE = 0.24$ ,  $p = .06$ : participants were more conservative when retention interval was short ( $M = -.02$ ,  $SEM = 0.04$ ) than when it was long ( $M = -.14$ ,  $SEM = 0.04$ ). There was no interaction between list type and retention interval ( $F < 1$ ,  $p = .43$ ).

For *B* items, there was no main effect of retention interval and no interaction ( $F_s < 1.56$ ,  $p_s > .22$ ). There was, however, a strong main effect of list type,  $F(2,88) = 16.15$ ,  $MSE = 0.07$ ,  $p < .001$ : responses were more liberal for *strong* lists than for both *short* and *long* lists and responses were more liberal for *long* lists than for *short* lists. Table 4.9 presents these results.

**Table 4.9. Bias across item types ( $c_a$ ) (Exp. 6).**

List type	A items			B items		
	<i>M</i>	( <i>N</i> = 46)	<i>SEM</i>	<i>M</i>	( <i>N</i> = 45)	<i>SEM</i>
<b>Short</b>	-.14	⌈ ⌈ n	0.05	-.05	⌈ ⌈ *	0.04
<b>Long</b>	-.14	⌈ ⊥ ***	0.04	-.14	⌈ ⊥ ***	0.04
<b>Strong</b>	0.05	⊥ ⊥	0.04	-.27	⊥ ⊥	0.03

*Note.*  $c_a$  = response bias (hits and false alarms at  $X_3$ , which separates *guess old* from *guess new* responses). A items = targets; B items = interference items (repeated in *strong* lists). Data collapsed across retention intervals. *n* non-significant; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .

### Experiment 5b vs. Experiment 6 (number of repetitions and item type)

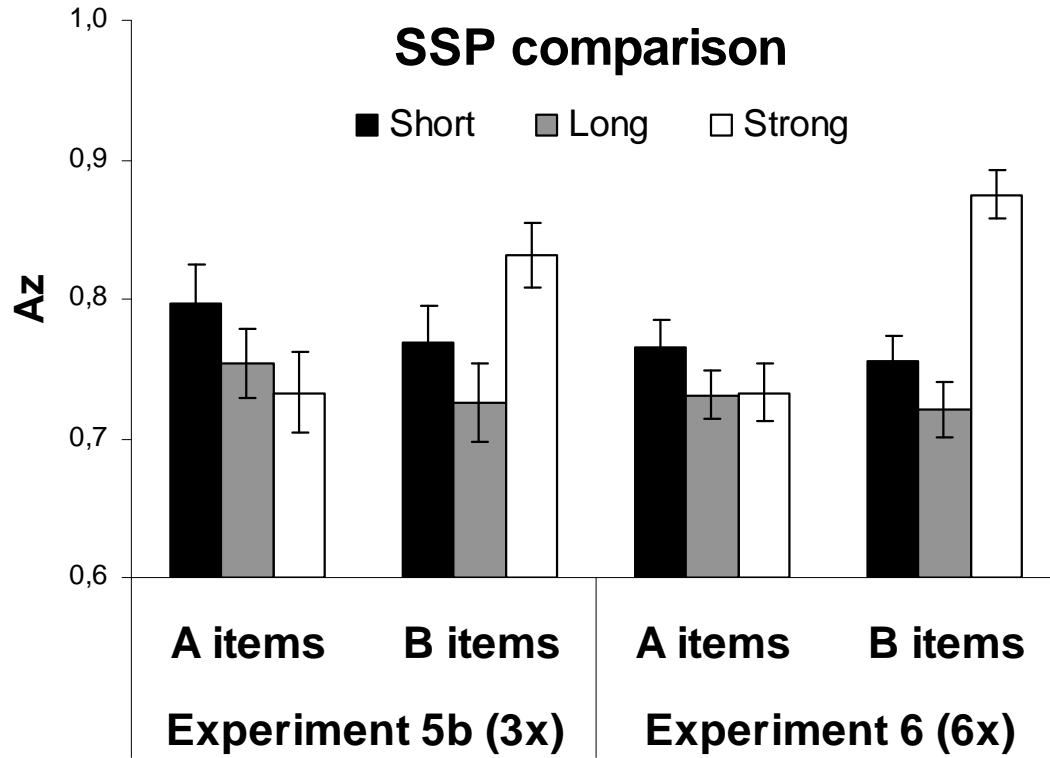
In Experiments 5b and 6, participants attended two experimental sessions on different days (3 study-test blocks in the same session). In Experiment 5b, *long* lists were twice as long as *short* lists and strong items in *strong* lists were shown 3 times. In Experiment 6, *long* lists were 3.5 times longer than *short* lists and strong items were presented 6 times. LLE and LSE were found in both experiments but in neither experiment did retention interval significantly modulate the size of the interference effects. Figure 4.7 illustrates the results with data collapsed across intervals.

A 2 [experiment: 5*b* (3x), 6 (6x)]  $\times$  3 (list type: *short*, *long*, *strong*)  $\times$  2 (retention interval: *short*, *long*) mixed-design ANOVA on sensitivity  $A_z$  revealed only a main effect of list type,  $F(2,132) = 7.27$ ,  $MSE = 0.01$ ,  $p = .001$ . Post-hoc LSD tests confirmed that sensitivity in *long* and *strong* lists was worse than in *short* lists, and that sensitivity in *long* and *strong* lists did not differ ( $p = .52$ ). All other main effects and interactions were not significant (all  $F$ s  $< 1.3$ ,  $ps > .26$ ). The results show no sign of modulation of LLE and LSE by manipulation strength. In other words, increasing the list length from 120 to 210 words did not change the size of the LLE. Similarly, increasing the presentation of *interference* items from 3 to 6 times did not change the magnitude of the LSE.

We also carried out a 2 [experiment: 5*b* (3x), 6 (6x)]  $\times$  2 (item type: *A*, *B*)  $\times$  2 (retention interval: *short*, *long*) mixed-design ANOVA on  $A_z$  for *strong* lists to assess the impact of number of repetitions on discriminability. The three-way ANOVA revealed a main effect of item type,  $F(1,62) = 68.02$ ,  $MSE = 0.01$ ,  $p < .001$ , showing that discrimination between *targets* and *SP lures* was better for *B* items (studied multiple times) than for *A* items (studied only once). There was also an interaction between experiment and item type,  $F(1,62) = 4.17$ ,  $MSE = 0.01$ , indicating that the difference in discriminability for *A* and *B* items was larger in Experiment 6, where strong items were presented 6 times ( $M_A = .74$ ,  $SEM = .02$ ;  $M_B = .88$ ,  $SEM = .02$ ), than in Experiment 5*b*, where they were presented 3 times ( $M_A = .74$ ,  $SEM = .03$ ;  $M_B = .83$ ,  $SEM = .02$ ). All other main effects and interactions were not significant ( $F$ s  $< 1.7$ ,  $ps > .19$ ). The results confirm that, not only the strength manipulation was effective in both experiments, but also that increasing the number of repetitions from 3 to 6 yielded the expected increase in sensitivity. Thus, the lack of modulation of the LSE by the number of presentations (3x vs. 6x) cannot be attributed to an ineffective strength manipulation.

To assess whether or not output interference was affecting performance, we conducted a 2 [experiment: 5*b* (3x), 6 (6x)]  $\times$  2 (item type: *A*, *B*)  $\times$  2 (retention interval: *short*, *long*) mixed-design ANOVA on  $A_z$  for *long* lists. The results revealed that, indeed, discrimination was lower for *B* items ( $M_B = .72$ ,  $SEM = .02$ ) compared to *A* items ( $M_A = .75$ ,  $SEM = .02$ ;  $F(1,62) = 4.14$ ,  $MSE = 0.01$ ) even though *A* and *B* items were indistinguishable to participants at study and even

though *A* and *B* items were exposed to the same amount of interference from subsequent items on the study list. The only difference between those item types was their relative position on the test list: a mixture of *A targets* and *A SP lures* was tested in the first half of the test list followed by a mixture of *B targets* and *B SP lures*. The result shows that output interference played a role in *B*-item sensitivity.<sup>5</sup>



**Figure 4.7. Sensitivity across repetitions and item types (Exps. 5b / 6).**

Sensitivity ( $A_z$ ) for *A* items in *long* and *strong* lists decreased relative to sensitivity in *short* lists (LLE and LSE, respectively) in both Experiments 5b and 6. *A* items were studied first (within first 60 study trials) and tested first (first 30 test trials). *B* items were studied first (within first 60 study trials intermixed with *A* items) and tested second (last 30 test trials). For *short* and *long* lists, sensitivity for *B* items was lower than for *A* items, consistent with output interference effects. Sensitivity for *B* items in *strong* lists was higher than in *short* and *long* lists because *B* items in *strong* lists were presented multiple times at study. Sensitivity was higher for *B* items in Experiment 6, where they were presented 6 times, compared to Experiment 5b, where they were presented 3 times. The result confirms that the increase in strength across experiments was effective. Despite that, the LSE did not increase across experiments. The LLE observed with *A* items was replicated with *B* items, attesting to the robustness of the effect. Although the study list was longer in Experiment 6 than in Experiment 5b (rising from 2:1 to 3.5:1 in terms of long-to-short list length ratio), the LLE did not increase. SSP = studied vs switched-plurality comparison;  $A_z$  = sensitivity; Error bars = SEM;  $N = 68$ .

Finally, we compared sensitivity of *A* items in *short* lists across experiments to evaluate whether the strength of manipulation (longer or stronger lists) was having

<sup>5</sup> *Output interference* is defined here as a decrease in sensitivity as a function of the position of an item in a test sequence, regardless of its original position in the study sequence (cf. Schulman, 1974).



any effect on the baseline condition. A 2 [experiment: 5*b* (3x), 6 (6x)]  $\times$  2 (retention interval: *short*, *long*) mixed-design ANOVA on  $A_z$  for *short* lists revealed no main effect of experiment and no main effect of retention interval ( $F_s < 1$ ,  $p_s > .36$ ). Thus, overall, there was no difference in the baseline condition across experiments. The interaction between experiment and retention interval, however, although not significant [ $F(1,66) = 2.59$ ,  $MSE = 0.01$ ,  $p = .11$ ], seemed large enough to warrant simple comparisons. Indeed, independent sample *t*-tests showed that, whereas sensitivity did not vary across experiments when retention interval was short [ $t(66) < 0.01$ ,  $p = .99$ ], it did approach significance when the interval was long [ $t(66) = 1.88$ ,  $p = .07$ ; Exp. 5*b*:  $M_A = .81$ ,  $SEM = .02$ ; Exp. 6:  $M_A = .75$ ,  $SEM = .03$ ]. Thus, retention interval appears to have slightly affected the baseline across experiments, as sensitivity was worse when study-test lag was longer.

#### 4.4.4. Discussion

The results of Experiment 6 largely replicate the results of Experiment 5*b*. In both experiments, reliable LLEs and LSEs were observed. In neither experiment has retention interval significantly affected the magnitude of the interference effects. The fact that both list-length and list-strength effects were found in experiments whose designs largely differed (e.g., Exp. 4 in Chapter 2 vs. Exp. 6 in this chapter) attests to the robustness of those effects. Also, the fact that retention interval here failed to modulate sensitivity is in agreement with the results from Experiments 4 and 5*b*. Importantly, retention interval did affect response bias in this experiment: participants were more conservative (i.e., responded “old” less often) when retention interval was short (10 s) than when it was long (120 s). The impact of retention interval on bias suggests that participants were sensitive to the interval manipulation. Yet, the manipulation failed to affect target-lure discriminability.

Another variable that failed to modulate interference was manipulation strength. Increasing list length from 120 words (Exp. 5*b*) to 210 words (Exp. 6) did not increase the size of the LLE. Similarly, increasing item strength from 3 (Exp. 5*b*) to 6 presentations (Exp. 6) did not change the size of the LSE, despite a reliable rise in discriminability for strong items across experiments. Effect sizes were instead lower than in Experiment 5*b* [LLE:  $d_z = 0.27$  vs. 0.40; LSE:  $d_z = 0.23$  vs. 0.74]. The fall in

effect size may be attributed to a non-significant decline in the baseline condition. Because performance for *short* lists fell slightly but performance for *long* and *strong* lists did not change, the net effect was a reduction in overall effect size. It is unclear why baseline changed without concurrent changes in the other two list types, since uncontrolled variables, such as longer study-test lags or higher levels of fatigue, would be expected to affect all three list types in a similar way. In addition, floor effects for *long* and *strong* lists are unlikely, as sensitivity revolved around .75. The present results differ from the results of Experiments 3 (2:1, 3x) and 4 (3.5:1, 6x; SSP comparison) in that LLE (but not LSE) did change in the longer list condition.

The null modulation of LLE by length, although surprising, is not unheard of. Cary and Reder (2003, Exps. 1 and 2) reported a similar result when list-length ratios were 2:1 and 3:1. They had participants studying lists of 16, 32, 48 and 64 items. Although the overall ANOVA across list lengths was significant, there was hardly any difference in sensitivity ( $d'$ ) between 32-item and 48-item lists. Sensitivity for 64-item lists, on the other hand, was distinctively lower (see Table 4.10; pairwise comparisons were not reported in the study). The small difference in sensitivity between 32-item and 48-item lists is even more surprising given that Cary and Reder's (2003) design contained some of the features highlighted by Dennis and Humphreys (2001) as possible confounds, such as longer test sequences for longer lists (length confounded with output interference), proactive design (length confounded with lower levels of attention to targets in longer lists) and distribution of targets throughout longer lists (length confounded with average study-test lags).

**Table 4.10. Sensitivity ( $d'$ ) in Cary and Reder (2003).**

Experiment	List length			
	16	32	48	64
<b>1 (RI = 0 s)</b>	2.66	2.32	2.29	2.02
<b>2 (RI = 300 s)</b>	1.95	1.51	1.59	1.23

*Note.* RI = retention interval. Proactive design in both experiments.

The increase in list-length ratio from 2:1 to 3:1 in Cary and Reder's (2003, Exp. 1 and 2) study, like in Experiment 6 here, barely changed their LLE. In addition, their retention interval manipulation only affected overall sensitivity (lower  $d'$  in long interval condition) but did not affect the list-length effect. Cary and Reder's (2003) results suggest that the null modulation of LLE here may have been caused by an

insufficient manipulation of length. This impression is reinforced by noting that mean sensitivity in long lists decreased (non-significantly) across our experiments ( $M_{Exp.5b} = .75$ ,  $SEM = .03$ ;  $M_{Exp.6} = .73$ ,  $SEM = .02$ ). Because of that, we refrain from making strong claims about the theoretical implications of the null LLE modulation.

The null LSE modulation observed here has also been previously observed. Diana and Reder (2005) found little difference in the magnitude of their (small) LSE when strong-item presentation increased from 6 (Exp. 1) to 11 times (Exp. 2). Diana and Reder (2005) used only *unrelated lures*, whereas we used only related lures. This suggests that lure type is not the critical factor underlying the null result here. Norman (1999, Exps. 4 and 4a) did find a larger LSE with *unrelated lures* when presentations rose from 3 to 6 times. Unlike our Experiments 5b and 6, Norman (1999) controlled for study-test lag across his Experiments 4 and 4a, resulting in similar performance for the baseline conditions across experiments. Any changes in effect size could thus be attributed to changes in performance in *strong* lists.<sup>6</sup>

**Table 4.11. Hits and false alarms in Norman (1999).**

Experiment	List strength			
	Short		Strong	
	HR	FAR	HR	FAR
<b>4 (6x, unr.)</b>	.91	.12	.66	.03
<b>4a (3x, unr.)</b>	.92	.13	.82	.07

*Note.* HR = hit rates; FAR = false-alarm rates; unr. = unrelated lures.

Table 4.11 summarises Norman's (1999, Exps. 4 and 4a) results. Although the data clearly shows a larger decrease in hits across list types in his Experiment 4 (6x) than in Experiment 4a (3x), the decrease in false alarms across list types, normally found in strength manipulations (e.g., Hirshman, 1995), was much less pronounced across experiments. In particular, false alarms for *strong* lists were very low in Norman's (1999) Experiment 4, suggesting a floor effect. If a floor effect for false alarms indeed occurred, then the LSE observed in Experiment 4 may have been overestimated. Moreover, because retention interval for *strong* lists in Experiment 4a was longer than in Experiment 4 and because longer intervals may reduce the size of LSEs (as found in our Experiments 2 and 3), the LSE in Experiment 4a may

<sup>6</sup> The average interval between studying an item and being tested on that item was 2.5 min. longer in Experiment 6 than in Experiment 5b. Study-test lag was not controlled here because our goal was to study retention interval, which would have changed had we kept the same lags across experiments.

have been underestimated. Taken together, these factors may have contributed to Norman's (1999) finding of a modulatory effect of number of repetitions on the LSE. The effect of repetitions on LSE, if it exists, may in fact be much smaller.

Another possible reason why increasing the number of presentations of strong items may not necessarily increase the size of the LSE is that participants tend to fail to encode additional features of the stimuli after their initial presentations. In fact, repeating an item (*banana*) more than one or two times does not normally improve participants' ability to discriminate the item from its switched-plurality lure (Hintzman et al., 1992). Participants are as good at rejecting lure *bananas* when *banana* is presented 3 times at study as they are when *banana* is presented 15 times at study. The obvious explanation – that participants are not paying attention to the repeated items – is ruled out by the fact that participants are reasonably accurate at determining how many times the repeated items have been presented. Thus, although participants store some information from repeated presentations of an item, they tend to ignore the details of that item if not stored during the item's first presentation. This phenomenon, dubbed *registration without learning*, has been replicated with pictures (Hintzman et al., 1992, Exp. 2) and auditory stimuli (Sheffert & Shiffrin, 2003) and is very resistant to instructional manipulations (Hintzman & Curran, 1995). To the extent that registration without learning occurred in Experiment 6, extra repetitions of strong items would have little impact on the memorability of weak items, thereby causing little change on the LSE. This possibility, however, seems unlikely. That is because participants here did profit from extra presentations, as discrimination for strong items in Experiment 6 was higher than in Experiment 5*b* (see Figure 4.7), indicating that participants encoded discriminative features of *targets* to a larger extent when they were presented more times. Thus, the null modulation of LSE observed here is probably not due to registration without learning.

One factor that probably reduced the impact of number of repetitions on the LSE was the repetition schedule. It is known that spaced repetition – repeating an item after long lags – produces better learning than massed repetition – repeating an item after short lags (see Dempster, 1996, for a review). Although repetitions were not massed in our experiments, they were not strictly spaced either, since the repetition

schedule adopted here had no restrictions (e.g., an item could be repeated immediately after its second presentation). Thus, although in the majority of cases repetitions occurred after more than one intervening item, there were cases in which spaced repetitions were in effect massed repetitions. Clearly, this feature of the design did not prevent strong items from being strengthened. Nonetheless, the repetition schedule used here may have contributed to a suboptimal manipulation.

In fact, Malmberg and Shiffrin (2005) showed that, in free recall tasks, LSEs are obtained with spaced repetitions but not with massed repetitions. Thus, spaced repetitions not only lead to better learning of repeated items but also to more interference on non-repeated items in free recall. As pointed out by Malmberg and Shiffrin (2005), repetitions seem most effective when they occur after a study item has left working memory. Malmberg and Shiffrin's (2005) result suggests that implementing an expanding repetition schedule (i.e., repeating items at ever increasing lags) may lead to stronger recognition LSEs. Because the null LSE modulation here could thus have been caused by low manipulation strength, we refrain from drawing strong theoretical conclusions from these results.

Some memory models may have problems explaining our data. BCDMEM cannot account for the positive LLE and LSE found across retention intervals and experiments, since it predicts interference only when retention interval is short. Likewise, REM predicts neither LLE nor LSE in SSP comparisons, contradicting the positive effects found across experiments and retention intervals. CLS and SAC, on the other hand, predict both effects. CLS predicts LLEs and LSEs of similar sizes in accord with the data, whereas SAC predicts larger LLEs than LSEs (although SAC can possibly fit effects of similar size). The lack of LLE and LSE modulation by manipulation strength, although counter to CLS and SAC's predictions, could be accounted for by low manipulation strength.

CLS explains LLE and LSE in terms of interference on its recollection component: extra items or strong items reduce activation of distinctive features of the other items on the list; both familiarity and recollection are affected but only the recollection component shows a net decrease in discrimination because lures are unlikely to trigger recollection and, consequently, are at floor. Thus, models

implementing a recollection mechanism with similar properties could, in principle, account for LSEs. A version of REM with a recall mechanism has already been proposed (Malmberg, Holden et al., 2004) to account for the registration without learning phenomenon (in its original form, REM incorrectly predicts ever increasing SP-lure false alarms with increasing target repetition). The recall version of REM, however, has not been tested in the context of list-strength manipulations.

Two final features of the results are worth mentioning. The first is that the ROC curves in Experiments 5a, 5b and 6 were curvilinear. An unequal-variance SDT model was able to fit individual participants' data well: only 4 out of 720 models across experiments and conditions were rejected at the .05 level. Participants had to use some form of recall in order to produce above-chance responses, since *targets* and *SP lures* were equally familiar. Thus, the fact that an unequal-variance SDT could fit the data well suggests that the source of recall information may be continuous in nature rather than all-or-none. The second feature of the data worth mentioning is the lack of a complete within-list, strength-based, mirror effect across experiments: hit rates increased from weak (*A*) to strong (*B*) items in *strong* lists, but the false alarms did not significantly decrease across item types. The result is consistent with participants adopting the same response criteria throughout the test list, despite a change in difficulty (from hard to easy) partway through. We discuss in more depth the theoretical implications of these findings in Section 4.6.3.

#### 4.5. Experiment 7: Retention interval, lure type, with new, 6x

In this experiment, we undertook to test contrasting predictions made by BCDMEM (Dennis & Humphreys, 2001) and CLS (Norman & O'Reilly, 2003). Both models predict larger interference effects with shorter retention intervals. However, they make distinct predictions with respect to lure types. BCDMEM predicts equal interference effects at short retention intervals for both *unrelated* and *switched-plurality lures*. CLS, on the other hand, predicts stronger interference effects at short retention intervals for *SP lures* than for *unrelated lures*.

Interference should be the same across lures in BCDMEM because the model treats *unrelated* and *SP lures* in the same way: both lure types are stored as distinct, single nodes in the model's input layer.<sup>7</sup> Consequently, any manipulation affecting one lure type should similarly affect the other. Thus, if retention interval modulates LLE and LSE with *SP lures* it should also modulate those effects with *unrelated lures*. By contrast, interference should differ across lures in CLS because study items are encoded in a distributed network and, consequently, can vary in their similarity to the lures at test. While *unrelated lures* are likely to be judged in terms of familiarity (less affected by length and strength manipulations), *SP lures* can be judged in terms of both familiarity and recollection. Recollection, which is likely to be triggered due to *SP lures'* similarity to *targets* and which allows confident rejection of similar lures, is highly affected by length and strength manipulations (see 1.6.3 for a description of why this is the case). Thus interference should be more apparent during *SP lure* decisions than during *unrelated lure* decisions. More importantly, if retention interval modulates LLE and LSE, it should do so to a larger extent when lures are highly similar (*SP lures*) than when they are less similar (*unrelated lures*).

In this experiment, like in Experiment 6, *long* lists were 3.5 times longer than *short* lists and strong items were presented 6 times in *strong* lists. Moreover, retention interval was manipulated within participants. Unlike Experiment 6, however, *unrelated lures* here were also presented at test. The introduction of *unrelated lures* permits the comparison of interference effects across lure types and thus a direct comparison of BCDMEM and CLS predictions. Any result showing differential LLE or LSE between *unrelated* and *SP lures* would constitute evidence for CLS and against BCDMEM. Conversely, a result showing similar changes in LLE and LSE across lure types would be evidence for BCDMEM and against CLS.

---

<sup>7</sup> There are two possibilities for *SP lures* in BCDMEM: either they are represented in the same node as their corresponding *targets* (i.e., target *banana* and lure *bananas* share the same node) or they are represented in different nodes. Sharing a node is not an option because *target-lure* discriminability should then be at chance, a prediction not borne out by the data. Representing *SP lures* on different nodes, on the other hand, cannot explain why discrimination between *targets* and *SP lures*, although above chance, is nonetheless far from perfect ( $A_z$  for strong *B* items is .88 even after 6 presentations). The prediction described here assumes that *targets* and *SP lures* are represented as separated nodes.

#### 4.5.1. Methods

##### Participants

Ninety-six University of Warwick students (47 males; mean age = 21.9,  $SD = 6.5$ ) participated in the study. The experiment lasted 60 min. (two 30-min. sessions on different days) and participants were paid £6.

List type	Short retention interval		
	Study	Distractor	Test
Short	[AB]	257.5 s	[tA, spA, unr]
Long	[AB] [CDEFG]	10 s	[tA, spA, unr]
Strong	[AB] [BBBBB]	10 s	[tA, spA, unr]

	Long retention interval		
	Study	Distractor	Test
Short	[AB]	367.5 s	[tA, spA, unr]
Long	[AB] [CDEFG]	120 s	[tA, spA, unr]
Strong	[AB] [BBBBB]	120 s	[tA, spA, unr]

**Figure 4.8. Design of Experiment 7.**

A-G = matched 30-word groups; [X,Y] = groups X and Y are merged and word order randomised; tA = targets from group A; spA = switched-plurality lures from group A; unr = unrelated lures.

##### Materials

Stimuli were 450 nouns from the MRC Psycholinguistic Database: imageability = 5.71 [5.02-6.52]; concreteness = 5.77 [5.00-6.48]; familiarity = 5.09 [4.00-6.16]; Kučera-Francis frequency = 16.63 occurrences per million [0-99]; word length = 5.47 [3-10]. Thirty words were used as fillers and the remaining 420 words were assigned to 14 groups of 30 words, matched for word characteristics. Of the 14 word groups, 3 consisted of *targets*, 8 consisted of *interference* words and 3 consisted of *unrelated lures*. Distinct samples were produced for each participant.



## Design and Procedure

Figure 4.8 illustrates the experimental design. Design and Procedure were similar to Experiment 6 with two differences. First, only *A targets* were tested. Second, *unrelated lures* were added at test. Test lists consisted of 60 words (15 *targets*, 15 *SP lures* and 30 *unrelated lures*). Retention interval and list type were manipulated within participants and responses were self-paced.

### 4.5.2. Results

#### Hits and false alarms

A three-way [2 (word type: *target*, *lure*)  $\times$  2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*)], repeated-measures ANOVA on proportion of “old” responses revealed a main effect of word type,  $F(1,95) = 1310.23$ ,  $MSE = 0.06$ ,  $p < .001$ , such that “old” responses were given more often to *targets* than to *lures* (*SP lures* and *unrelated lures* were collapsed in the analysis). There was also a main effect of list type,  $F(2,190) = 42.30$ ,  $MSE = 0.01$ ,  $p < .001$ , such that the proportion of “old” responses in *strong* lists was lower compared to the proportion in *short* and *long* lists. In addition, word type and list type interacted,  $F(2,190) = 3.41$ ,  $MSE = 0.01$ , suggesting that the proportion of “old” responses for *targets* in *long* and *strong* lists was lower compared to responses for *targets* in *short* lists but that the proportion of “old” responses for *lures* moved in opposite directions in *long* and *strong* lists (false alarms increased in *long* lists and decreased in *strong* lists relative to *short* lists). Finally, retention interval interacted with list type,  $F(2,190) = 4.01$ ,  $MSE = 0.01$ , suggesting that the drop in the proportion of “old” responses in *strong* lists was larger when the retention interval was short. There was no main effect of retention interval, no interaction between retention interval and word type and no three-way interaction among all three variables,  $F_s < 1.6$ ,  $p_s > .20$ .

Separate two-way [2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*)], repeated-measures ANOVAs were carried out on hits, SP false alarms (to *SP lures*) and unrelated false alarms (to *unrelated lures*). For hits, there was an effect of list type,  $F(2,190) = 25.23$ ,  $MSE = 0.02$ ,  $p < .001$ , such that hits were lower for *strong* lists than for *short* and *long* lists. Hits did not reliably differ between

*short* and *long* lists ( $p = .23$ ). There was also an interaction between retention interval and list type,  $F(2,190) = 3.38$ ,  $MSE = 0.02$ , showing that the drop in *strong* lists relative to *short* and *long* lists was larger when the interval was short.

For SP false alarms, there was an effect of list type,  $F(2,190) = 11.89$ ,  $MSE = 0.02$ ,  $p < .001$ . Post-hoc LSE comparisons revealed that SP false alarms in *strong* lists were lower than in *short* and *long* lists ( $ps < .001$ ), whereas SP false alarms did not differ between *short* and *long* lists ( $p = .34$ ). There was no effect of retention interval and no interaction between retention interval and list type ( $Fs < 1$ ,  $ps > .35$ ).

Finally, for unrelated false alarms, there was an effect of list type,  $F(2,190) = 4.32$ ,  $MSE = 0.01$ ,  $p < .001$ ; LSD comparisons showed that unrelated false alarms in *strong* lists were lower than in *short* and *long* lists ( $ps < .001$ ) and that false alarms did not differ between *short* and *long* lists ( $p = .51$ ). There was also a main effect of retention interval,  $F(1,95) = 4.32$ ,  $MSE = 0.01$ , such that there were fewer unrelated false alarms when retention interval was short than when it was long. Retention interval and list type did not interact ( $p = .22$ ).

**Table 4.12. Hits and false alarms (Exp. 7).**

List type	HR Targets				FAR SP lures				FAR Unrelated			
	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>
<b>Short</b>	.75	τ	τ	.01	.40	τ	τ	.02	.14	τ	τ	.01
		n				n				n		
<b>Long</b>	.74	τ	⊥ ***	.01	.41	τ	⊥ ***	.02	.15	τ	⊥ ***	.01
		***				***				***		
<b>Strong</b>	.66	⊥	⊥	.02	.34	⊥	⊥	.02	.10	⊥	⊥	.01

*Note.* HR = hits; FAR = false alarms; SP = switched plurality; *n* non-significant; \*\*\*  $p < .001$ .  $N = 96$ .

Hits and false alarms, collapsed across retention intervals, are presented in Table 4.12. Hits and false alarms broken down by retention intervals are presented in Appendix 1, together with single-point measures of sensitivity ( $d'$ ) and bias ( $c$ ).

There was no effect of list length in the SU comparison, as the interaction between word type (*target*, *unrelated lure*) and list type (*short*, *long*) was not significant ( $p = .17$ ). Word type marginally interacted with retention interval ( $p = .08$ ), hinting that

hits increased and false alarms decreased from long to short retention intervals. In the SSP comparison, there was a marginal interaction between word type (*target*, *SP lure*) and list type (*short*, *long*),  $F(1,95) = 2.86$ ,  $MSE = 0.01$ ,  $p = .09$ . The interaction showed that hits decreased and false alarms increased from *short* to *long* lists. Thus, a marginal effect of list length was revealed in the SSP comparison across intervals.

There was an effect of strength in both SU and SSP comparisons: the interactions between word type (*target* vs. *SP lure* and *target* vs. *unrelated lure*) and list type (*short*, *strong*) were significant ( $F_s > 3.86$ ,  $p_s < .05$ ). The interactions indicate that the fall in hits from *short* to *strong* lists was larger than the fall in false alarms. In the SU comparison, there was also a main effect of retention interval,  $F(1,95) = 4.71$ ,  $MSE = 0.01$ , such that fewer “old” responses were given when retention interval was short. Retention interval also interacted with list type,  $F(1,95) = 4.34$ ,  $MSE = 0.01$ , suggesting that the drop in the proportion of “old” responses from *short* to *strong* lists was larger when retention interval was short. In the SSP comparison, by contrast, there was no main effect of retention interval,  $F < 1$ ,  $p = .80$ , although the interaction between retention interval and list type was marginally significant,  $F(1,95) = 3.23$ ,  $MSE = 0.02$ ,  $p = .08$ . As with the SU comparisons, the latter result suggests that the drop in proportion “old” responses from *short* to *strong* lists was larger when the retention interval was short.

In sum, the results from hits and false alarms revealed harmful effects of list length and list strength manipulations on memory in SSP comparisons. In SU comparisons, however, only strength manipulations affected performance.

### Sensitivity

A total of 864 unequal-variance Gaussian models were fitted to individual participants' confidence data (96 participants  $\times$  3 list types  $\times$  3 comparison types: SU, SSP and SPU). The data of 24 participants were excluded due to poor fits. The results refer to the parameter estimates of the remaining 72 participants. Table 4.13 summarises the results collapsed across retention intervals.

**Table 4.13. Sensitivity ( $A_z$ ) across retention intervals (Exp. 7).**

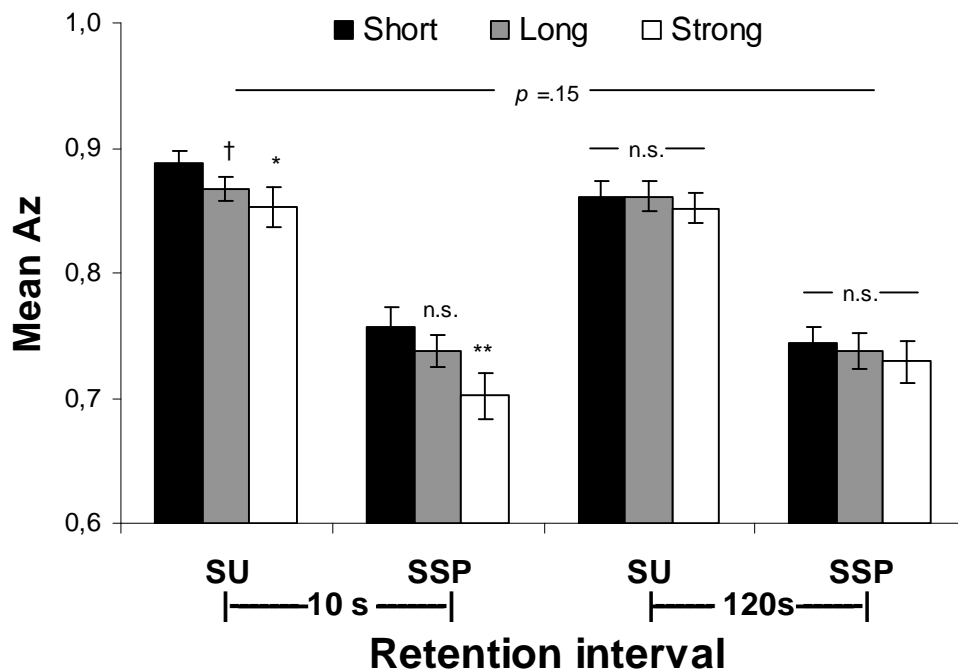
List type	SU			SSP			SPU				
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>		
Short	.88	$\tau$ $n$ $\perp$	$\tau$ $\dagger$	.01	.75	$\tau$ $n$ $\perp$	$\tau$ $n$ $\perp$	.01	.64	$\tau$ $n$ $\perp$	.02
Long	.86	$\tau$ $n$ $\perp$	$\perp$	.01	.74	$\tau$ $\dagger$ $\perp$	$\perp$	.01	.65	$\tau$ $n$ $\perp$	.02
Strong	.85	$\tau$ $n$ $\perp$	$\perp$	.01	.72	$\dagger$ $\perp$	$\perp$	.01	.69	$\tau$ $n$ $\perp$	.01

Note.  $A_z$  = area under the ROC; SU = studied vs. unrelated; SSP = studied vs. switched-plurality; SPU = switched-plurality vs. unrelated.  $n$  non-significant;  $\dagger p < .10$ ;  $* p < .05$ ;  $** p < .01$ .  $N = 72$ .

Separate 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*) repeated-measures ANOVAs were carried out on the sensitivity measure ( $A_z$ ) for each discrimination type (SU, SSP and SPU). In the SU comparison, there was no main effect of retention interval and list type and no interaction between the two variables,  $F_s < 2.16$ ,  $p_s > .12$ . In the SSP comparison, by contrast, there was a main effect of list type,  $F(2,142) = 3.90$ ,  $MSE = 0.01$ . Post-hoc LSD comparisons showed that participants were better at discriminating targets from lures in *short* lists than in *strong* lists (i.e., LSE,  $p = .01$ ). Sensitivity did not significantly differ between both *short* and *long* lists (i.e., null LLE,  $p = .33$ ) and *long* and *strong* lists ( $p = .09$ ). There was no main effect of retention interval and no interaction with list type,  $F_s < 1.7$ ,  $p_s > .19$ . In the SPU comparison, there was a main effect of list type,  $F(2,142) = 6.73$ ,  $MSE = 0.02$ ,  $p = .002$ , such that *pseudodiscrimination* was higher for *strong* lists than for *short* lists (i.e., negative LSE;  $p = .001$ ), higher for *strong* lists than for *long* lists ( $p = .02$ ) but similar for *short* and *long* lists ( $p = .24$ ). There was no main effect of retention interval and no interaction with list type,  $F_s < 1$ ,  $p_s > .54$ .

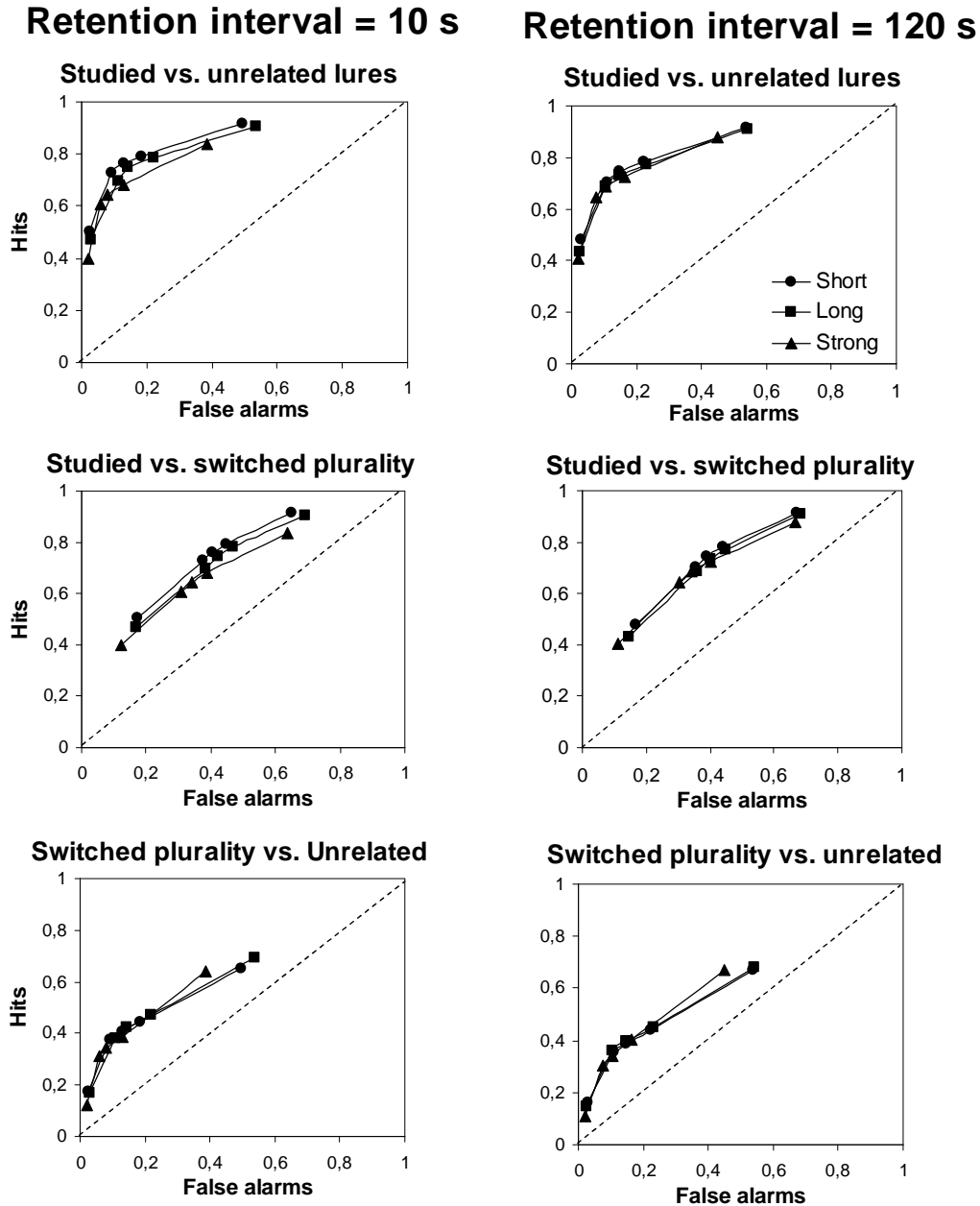
Although there was no significant interaction between retention interval and list type, there was a trend towards larger effects at short intervals. To better assess this trend, we conducted separate one-way ANOVAs on  $A_z$  for each discrimination type (SU, SSP and SPU) and retention interval (*short* vs. *long*) with list type as the independent variable. In the SU comparison, there was a main effect of list type at short intervals,  $F(2,142) = 3.26$ ,  $MSE = 0.01$ . Post-hoc LSD comparisons revealed lower discrimination for *long* and *strong* lists compared to *short* lists ( $p_s = .05$  and  $.02$ , respectively). Thus, at short retention intervals, both LLE and LSE were found in the SU comparison. By contrast, there was no effect of list type at long intervals ( $F < 1$ ,  $p = .76$ ). Similarly, in the SSP comparison, there was a main effect of list

type,  $F(2,142) = 5.18$ ,  $MSE = 0.01$ ,  $p < .01$ , at short but not at long intervals ( $F < 1$ ,  $p = .70$ ). Pairwise comparisons showed that  $A_z$  in the short interval condition was lower for *strong* lists than for both *short* and *long* lists ( $ps = .003$  and  $.045$ ) but  $A_z$  did not significantly differ between *short* and *long* lists ( $p = .26$ ). Thus, an LSE was found in the absence of an LLE in the SSP comparison when retention interval was short. There was no effect of list type at long intervals ( $F < 1$ ,  $p = .70$ ). Finally, in the SPU comparison, there was a main effect of list type at both short and long intervals ( $Fs = 3.41$  and  $3.59$ ,  $ps = .04$  and  $.03$ ), such that *pseudodiscrimination* was higher for *strong* lists than for *short* lists (i.e., negative LSE;  $ps = .007$  and  $.02$ ), higher for *strong* lists than for *long* lists ( $ps = .05$  and  $.08$ ) but similar for *short* and *long* lists ( $ps = .54$  and  $.36$ ). Overall, the analyses at each retention interval suggest that interference effects were somewhat larger when retention interval was short.



**Figure 4.9. Sensitivity across retention intervals (Exp. 7).**

Sensitivity for *long* lists was lower than for *short* lists only in SU comparisons at short retention intervals (10 s). Sensitivity for *strong* lists was lower than for *short* lists in both SU and SSP comparisons but only at short intervals. At long retention intervals (120 s), sensitivity did not change across list types. Significance values (†, \*, \*\*) refer to performance relative to *short* lists; the  $p$ -value at the top refers to the list type  $\times$  comparison type interaction term collapsed across retention intervals. SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality.  $A_z$  = sensitivity (area under ROC). Error bars = SEM. *n.s.* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ .  $N = 72$ .



**Figure 4.10.** ROC curves across retention intervals (Exp. 7).

For the short retention interval condition, the ROC curves for *long* and *strong* lists fall below the curve for *short* list in both “Studied vs. unrelated” and “Studied vs. switched plurality” comparisons and above it in the “Switched plurality vs. unrelated” comparison. For the long interval condition, the differences across list types were less pronounced (and not significant).  $N = 96$ .

A 3 (list type: *short*, *long*, *strong*)  $\times$  2 (discrimination type: SU, SSP)  $\times$  2 (retention interval: *short*, *long*) repeated-measures ANOVA on  $A_z$  was conducted to assess whether the impairment in sensitivity in *long* and *strong* lists was specific to SSP comparisons and whether it was modulated by retention interval. There was no interaction between list type and comparison type,  $F(1,71) = 2.12$ ,  $MSE = 0.01$ ,  $p = .15$ , suggesting that the interference effects did not differ dramatically between SU

and SSP comparisons. There was also no interaction between list type and retention interval,  $F(2,142) = 1.65$ ,  $MSE = 0.01$ ,  $p = .20$ , suggesting that retention interval did not reliably modulate the interference effects. However, the data trends here together with the one-way ANOVAs at each retention interval (previous paragraph) suggest that length and strength manipulations have a higher impact on sensitivity when retention interval is short. Figure 4.9 illustrates these results.

The differences in sensitivity across retention intervals can also be observed in the ROC curves. Figure 4.10 shows the ROC curves for each retention interval, discrimination type and list type. The curves for *long* and *strong* lists fall slightly below the curve for *short* lists in the SU comparison, more clearly below it in the SSP comparison and slightly above it in the SPU comparison. The differences across lists are clearer for the short retention interval condition.

### Bias

Separate 2 (retention interval: *short*, *long*)  $\times$  3 (list type: *short*, *long*, *strong*), repeated-measures ANOVAs on the bias measure ( $c_a$ ) were carried out for each discrimination type (SU, SSP and SPU). For the SU comparison, the ANOVA revealed a main effect of list type,  $F(2,142) = 36.19$ ,  $MSE = 0.08$ ,  $p < .001$ , such that participants were more conservative with *strong* lists than with *short* and *long* lists. Retention interval interacted with list type,  $F(2,142) = 6.67$ ,  $MSE = 0.06$ ,  $p = .002$ , showing that the rise in response bias was larger when retention interval was short. For the SSP comparison, there was also a main effect of list type,  $F(2,142) = 26.76$ ,  $MSE = 0.09$ ,  $p < .001$ , such that participants were more conservative with *strong* lists; there was also an interaction between retention interval and list type,  $F(2,142) = 4.28$ ,  $MSE = 0.08$ , indicating that the difference in response bias between *strong* lists and both *short* and *long* lists was larger at short retention intervals. For the SPU comparison, there was a main effect of list type,  $F(2,142) = 27.09$ ,  $MSE = 0.09$ ,  $p < .001$ , showing that participants were more conservative with *strong* lists. There was no main effect of retention interval,  $F < 1$ ,  $p = .67$ , but there was a marginal interaction between list type and retention interval,  $F(2,142) = 2.29$ ,  $MSE = 0.09$ ,  $p = .10$ , suggesting that the increase in bias in the SPU comparison was higher for *strong* lists when retention interval was short. Table 4.14 shows these results

collapsed across retention intervals (data broken down by retention intervals is presented in Appendix 1).

**Table 4.14. Bias ( $c_a$ ) across discrimination types (Exp. 7).**

List type	SU				SSP				SPU			
	$M$			$SEM$	$M$			$SEM$	$M$			$SEM$
<b>Short</b>	0.19	$\top$	$\top$	0.03	-.22	$\top$	$\top$	0.03	0.67	$\top$	$\top$	0.04
		$n$				$n$				$n$		
<b>Long</b>	0.22	$\perp$	***	0.03	-.18	$\perp$	***	0.03	0.66	$\perp$	***	0.04
		$\top$				$\top$				$\top$		
<b>Strong</b>	0.45	***	$\perp$	0.03	0.02	***	$\perp$	0.04	0.89	***	$\perp$	0.03
		$\perp$				$\perp$				$\perp$		

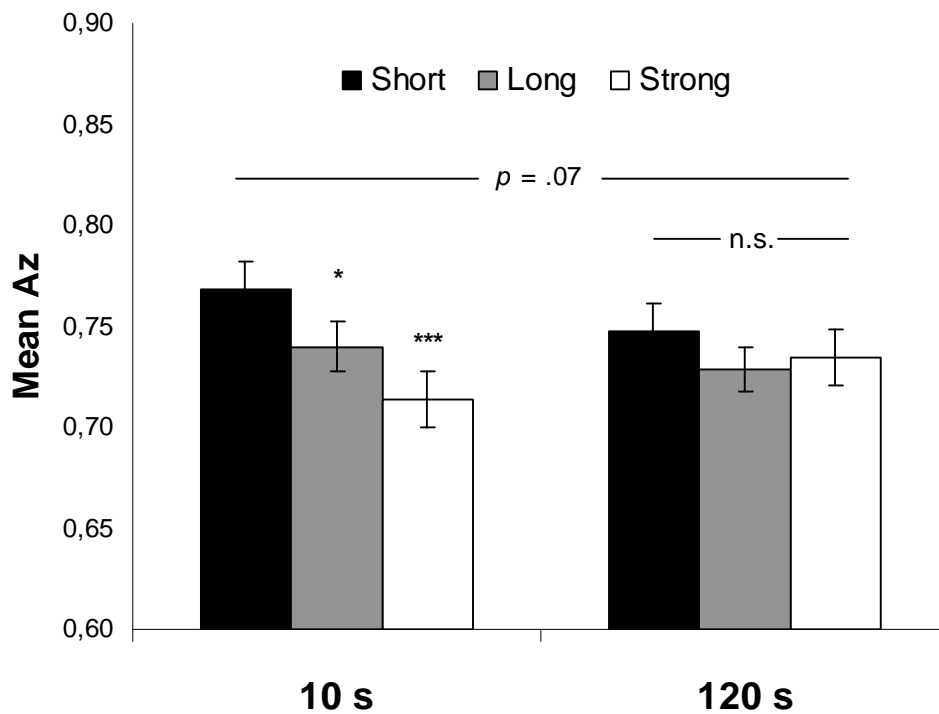
*Note.*  $c_a$  = response bias (from  $X_3$ ); SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality lures; SPU = switched-plurality vs. unrelated.  $n$  non-significant; \*\*\*  $p < .001$ .

#### Experiment 6 vs. Experiment 7 (without-new vs. with-new lures)

In Experiment 6, the test lists contained *targets* or *SP lures* (without-new condition), whereas in Experiment 7, they also included *unrelated lures* (with-new condition). It was hypothesised that discriminability would be better in the without-new condition because participants would be presumably less encouraged to rely on error-prone familiarity and more likely to engage in recollection to base their recognition decisions (Heathcote et al., 2006). To assess whether the presence of *unrelated lures* can affect discriminability, performance on *A* items in Experiment 6 was compared to performance on SSP discrimination in Experiment 7.

The 2 [experiment: 6 (*without-new*), 7 (*with-new*)]  $\times$  3 (list type: *short*, *long*, *strong*)  $\times$  2 (retention interval: *short*, *long*) mixed-design ANOVA on  $A_z$  yielded no main effect of experiment,  $F < 1$ ,  $p = .63$ . Thus, the presence of *unrelated lures* caused no detectable change in sensitivity across experiments (although  $A_z$  varied in the predicted direction; without-new:  $M = .74$ ,  $SEM = .02$ ; with-new:  $M = .73$ ,  $SEM = .02$ ). The ANOVA, however, revealed a main effect of list type,  $F(2,232) = 5.82$ ,  $MSE = 0.01$ ,  $p = .003$ ; LSD tests confirmed that sensitivity in *long* and *strong* lists was worse than in *short* lists ( $ps = .02$  and  $.001$ , respectively), and that sensitivity in *long* and *strong* lists did not differ ( $p = .33$ ). Finally, retention interval marginally interacted with list type,  $F(2,232) = 2.76$ ,  $MSE = 0.05$ ,  $p = .07$ . One-way ANOVAs confirmed the trend revealed by the interaction: there was an LLE and an LSE at short retention intervals ( $ps = .04$  and  $.001$ , respectively) but not at long intervals ( $ps = .16$  and  $.28$ , respectively). Figure 4.11 illustrates this interaction.





**Figure 4.11. Sensitivity across retention intervals (Exps. 6-A / 7-SSP).**

Test lists contained *SP lures* and *unrelated lures* in Experiment 7 but only *SP lures* in Experiment 6. List type interacted with retention interval such that LLE and LSE were significant at short retention intervals (10 s) but not at long intervals (120 s). The presence of *unrelated lures* at test did not affect performance (no effect of experiment). Data collapsed across experiments (Exp. 6: A items only; Exp 7: studied vs. switched-plurality comparison only). Error bars = SEM. *n.s.* non-significant; \*  $p < .05$ ; \*\*\*  $p = .001$ ;  $p$ -value on top bar represents list type  $\times$  retention interval interaction term.  $N = 118$ .

#### 4.5.3. Discussion

In Experiment 7, an LLE was found in the SU comparison only when retention interval was short. Similarly, an LSE was found in both SU and SSP comparisons only when retention interval was short. By contrast, a negative LSE was found in SPU comparisons at both short and long retention intervals.

Before discussing the theoretical implications of these data, there are two important qualifications to make. First, the LSE in the SU comparison may have been overestimated due to floor effects on false alarms. The proportion of *unrelated lure* false alarms that had to be corrected to avoid infinite  $d'$  values was much higher in *strong* lists (24% at short retention interval and 16% at long interval) than in *short* (7% and 8%) and *long* (9% and 6%) lists.<sup>8</sup> False alarms could have fallen further, as

<sup>8</sup> A complete list of raw data corrections across experiments is provided in Appendix 3.

a consequence of criterion shifts in *strong* lists, had they not reached floor. Thus, the difference between hits and false alarms in *strong* lists in the SU comparison may have artifactually shrunk. Second, the null LLE in the SSP comparison may have been underestimated due to inter-list interference. When only the data from the first study-test cycle was analysed, the LLE in the SSP comparison was significant at both short ( $p = .02$ ) and long ( $p = .05$ ) retention intervals, despite the loss of power incurred by the switch to a between-participant comparison. Because these particular data (LSE in SU and null LLE in SSP) may have been confounded with other factors, their theoretical implications will be played down.

Table 4.15 presents LLE and LSE effect sizes across comparison types either with data analysed across study-test blocks (3 blocks per session; list type manipulated within participants) or with data analysed only from the first study-test block (1 block per session; list type manipulated between participants). The table shows two major trends: effect sizes are larger when retention interval is short and effect sizes are larger in SSP comparisons than in SU comparisons.

**Table 4.15. Effect sizes of LLEs and LSEs (Exp. 7).**

Effect	All study-test blocks (matched pairs)				First study-test block (independent samples)			
	SU		SSP		SU		SSP	
	10 s	120 s	10 s	120 s	10 s	120 s	10 s	120 s
<b>LLE</b>	<b>0.26</b>	0.00	<u>0.16</u>	<u>0.04</u>	0.33	0.28	<b>0.41</b>	0.33
<b>LSE</b>	<u><b>0.29</b></u>	<u>0.11</u>	<b>0.39</b>	0.11	<u>0.12</u>	<u>0.21</u>	0.20	0.12

*Note.* SU = studied vs. unrelated lure; SSP = studied vs. switched-plurality lure. Retention interval conditions (10 s and 120 s). **Bold** = significant effects; underline = possible role of confounds.

Overall, the results provide mixed evidence for the BCDMEM model (Dennis & Humphreys, 2001). At first blush, the data seem to confirm BCDMEM's prediction that LLE and LSE should be apparent only at short intervals in both SU and SSP comparisons (see Figure 4.9). The effects should be similar at both comparisons because *unrelated* and *SP lures* are treated similarly by the model. However, given that the LSE in the SU comparison may be artifactual, the resulting pattern – LSE larger in SSP than in SU comparison – suggests different effects of retention interval across lure types, contrary to the model's prediction.

The CLS model (Norman & O'Reilly, 2003), on the other hand, predicts both trends. That is, the model predicts that effect sizes should be larger at shorter retention intervals and larger in SSP than in SU comparisons. The one caveat here is that CLS also predicts small but significant LLE and LSE in the SSP comparison at the long interval condition. In fact, Norman (2002, Exp. 2) reported the first LSE in recognition on that very experimental condition (120 s interval). Experiment 7 was based on Norman's (2002, Exp. 2) design and yet our results did not replicate his finding. One possible reason for the discrepancy is inter-list interference. In his doctoral thesis, Norman (1999, Exp. 2) found an LSE in the first two study-test blocks of his Experiment 2, but not in the last two blocks, suggesting that inter-list interference attenuates the size of list-strength effects. The discrepancy between our result and his could thus be similarly accounted for if the inter-list interference in our Experiment 7 was higher than in Norman's (2002) Experiment 2 (perhaps as a result of our added list-length manipulation).

Inter-list interference, however, is unlikely to have played a major role here because LSE effect sizes in the long interval condition barely changed when data from all study-test blocks was compared to data from the first block ( $d_z$ s = 0.11 vs. 0.12, respectively). The discrepant results cannot be attributed to lack of power either as the probability of detecting an LSE as large as the one reported by Norman (2002, Exp. 2) was .96. Although an LSE was not found at long intervals, the fact that effect sizes increased across intervals to a greater extent in SSP comparisons is consistent with CLS's prediction that the hippocampal model (recollection) should be more impaired than the cortical model (familiarity) in situations where study words are dissimilar.

The results of Experiment 7 alone provide only weak evidence for an effect of retention interval on LLE and LSE. But the comparison between Experiments 6 (A items) and 7 (SSP discrimination), with its increased power, clearly showed that retention interval can modulate the interference effects (see Figure 4.11). The relative size of that modulation, however, was smaller than anticipated.

## 4.6. Discussion of Experiments 5 to 7

### 4.6.1. Empirical summary

The main goal of Experiments 5 to 7 was to address some of the shortcomings of Experiments 1 to 4. The latter were largely modelled after Dennis and Humphreys (2001), whereas the former followed the design introduced by Norman (2002): study times were shortened and kept constant across list manipulations and strong items were included on the test list. Moreover, we extended Norman's (2002) design by adding list-length and retention interval manipulations. We also tested whether changes in manipulation strength (longer lists and stronger items) were followed by changes in interference effects. Finally, we tested whether the number of study-test blocks per session and whether the presence of *unrelated lures* at test would modulate the sizes of the interference effects. In general, the results of Experiments 5 to 7 confirmed and extended the results of Experiment 1 to 4.

First, the fact that LSEs were found in Experiments 5*b*, 6 and 7 rules out the possibility that the effects observed in Chapter 3 were due to the shorter study-test lags in the *strong* list conditions (recall that encoding was self-paced in Experiments 1 to 4, leading to shorter lags for *strong* lists). Second, the fact that LLEs and LSEs were modulated by retention interval in Experiments 6 and 7 (analysed together) argues against the possibility that the effect in Chapter 3 was caused by qualitative differences between the interval conditions (short intervals in Chapter 3 had no videogame phase between study and test, whereas a 10-s interval separated study and test in this chapter 3). Third, the idea that LSE (and to a lesser extent LLE) is stronger under conditions requiring recollection was supported by a trend in Experiment 7 towards stronger effects in the recollection-dependent SSP comparison than in the SU comparison (the trend could be significant if the LSE in the SU comparison is an artifact).

The lack of modulation by manipulation strength in Experiments 3 and 4 was also observed in Experiments 5*b* and 6. Manipulation strength was varied by increasing length ratio from 2:1 to 3.5:1 and strong-item presentation from 3 to 6 times. The fact that no modulation occurred in both pairs of experiments could be interpreted as

evidence that an upper bound has been reached in the magnitude of the effects. Alternatively, the null results might indicate that stronger manipulations need considerably different experimental designs to enable the detection of larger effects. Further increasing length and strength with the current design may simply cause participants to lose focus, with little gain in terms of interference.

For list-length manipulations, Dennis and Humphreys (2001, Exp. 1) suggested breaking up the study list into smaller chunks, interspersed with an interesting distractor activity (e.g., a puzzle task), in the hope of reducing attention loss. Higher levels of attention at study would be helpful in detecting larger LLEs because interference items would presumably be better encoded and, possibly, more likely to affect performance of other items on the list. Alternatively, if LLEs are simply the by-product of confounds such as differential attention loss in longer lists, as argued by Dennis and Humphreys (2001), then the increased attention levels afforded by that new design should either keep LLEs at the same level or make them disappear completely. In fact, Dennis and Humphreys (2001) reported the latter result in their Experiment 1 when using this design.

For list-strength manipulations, it has been known that simply repeating the same item in the same context may not be sufficient to enable complete encoding of the item's features, as attested by the registration-without-learning phenomenon (Hintzman et al., 1992). Complete encoding (i.e., noticing that the study item is *bananas*, not *banana*, and storing the correct version) requires overt response and feedback at study (Hintzman & Curran, 1995, Exp. 4). Thus, providing feedback at study may encourage stronger and more accurate encoding of study items, possibly improving the chances of interference towards weak items on the list (as predicted by the CLS model). Conversely, stronger and more accurate encoding of study items may lead them to become even more differentiated, reducing rather than increasing interference effects (as predicted by the REM model).

The number of experimental sessions modulated the interference effects. Participants underwent either 6 study-test blocks in one session or 3 blocks in two sessions. In the former, no interference effects were found; in the latter, both LLE and LSE were found. The effects were largely due to a change in the baseline

condition: discrimination in *short* lists was better with two sessions. The result suggests that interference effects may be masked when participants carry out several study-test cycles. Indeed, Diana and Reder (2005, Exp. 1) noticed that when participants underwent two blocks in one session, performance in the second block was lower regardless of list type (*short* vs. *strong*), prompting them to analyse only first-block data. We also analysed first-block data across experiments; the result patterns did not change except in Experiment 7, where the absent LLE in the SSP comparison emerged only when data from the first-block was analysed. The reason behind these modulatory effects is unclear. The hypothesis that longer sessions would mask changes in the variance of the underlying familiarity distribution, thereby reducing length and strength effects, was not supported. Instead, the SDT estimates showed smaller changes in mean familiarity values in longer sessions. Nonetheless, the results indicate that future research on interference effects should strive to keep the number of study-test cycles to a minimum.

Retention interval appeared to modulate LLE and LSE. The unreliability of the effect, however, is in a sense surprising. Intuitively, one would predict an increase in interference when the interfering items are stronger. Interfering items should be stronger in the short interval conditions because they have been recently presented. At longer delays, there is time for the activation of interfering items to decay. This account is not only intuitive but has also been explicitly implemented in a version of the CLS model in which neural network weights were allowed to vary with time (Norman & O'Reilly, 2003, p. 632). Yet, the empirical evidence is not conclusive. The between-experiment comparisons in Chapter 3 (Exps. 2 and 3) and in this chapter (Exps. 6 and 7) suggest a real effect. And in Experiment 7 there was a clear (non-significant) trend towards larger interference effects at shorter intervals. However, the effect was neither replicated in Experiment 4, where retention interval was manipulated between participants, nor was it replicated in Experiments 5(a,b) and 6, where retention interval was manipulated within participants.

In order to establish whether or not retention interval really modulates interference effects it may be necessary to widen the range of intervals from seconds to hours. This range is justified by research showing that it takes delays of over 40 min. for participants to adjust their response criteria on a trial-by-trial basis in a way

consistent with the relative recency of the studied items (Singer & Wixted, 2006). In other words, the difference in strength between older and recent items becomes apparent only after long delays, suggesting that the modulatory role of retention interval on LLE and LSE may also become apparent only after long delays.

The LSE in Experiment 5*b*, where only *targets* and *SP lures* were tested was much higher than in Experiment 3, where *unrelated lures* were also tested. The result is indicative that having only *SP lures* at test may increase the size of the interference effects. However, the comparison between Experiments 6 (without *unrelated lures*) and 7 (with *unrelated lures*), which is more meaningful given the similarity of their designs, showed no clear difference. Moreover, there was no overall improvement on sensitivity in the without-new condition compared to the with-new condition. Thus, we were not able to replicate Heathcote et al.'s (2006, Exps. 2 and 4) finding that discriminability is improved when only *SP lures* are tested.

To summarise, the results of Experiments 5 to 7 show that LLE and LSE are real effects, as they were consistently obtained across experiments. Both LLE and LSE were modulated by the relative contribution of recollection at test, by the number of study-test blocks in an experimental session and (weakly) by retention interval. Manipulation strength, however, had no reliable impact on effect sizes.

#### 4.6.2. Relation to other experiments

The list-length effects reported here replicate the LLE found by Cary and Reder (2003, Exp. 3). This is relevant because theirs was the first study to show an LLE after controlling for all the confounds identified by Dennis and Humphreys (2001). Thus, the fact that we consistently found the effect reinforces the case for the existence of LLEs in recognition. The results here also extend Cary and Reder's (2003) by showing a modulatory effect of retention interval, whereby LLEs were slightly larger when retention interval was short. In addition, our data showed a modulatory effect of number of study-test blocks: LLE was reduced when 3 study-test blocks were conducted in a session, compared to one block per session (see Table 4.15), and disappeared with 6 blocks per session (Experiment 5*a*).

The list-strength effects found in Experiments 5*b*, 6 and 7 replicate the LSE reported by Norman (2002, Exp. 2). His result was important because it was the first time an LSE was observed in a recognition memory task. The LSE in Experiment 5*b* in particular provides strong evidence for the existence of list-strength effects in recognition, since the effect was observed without a concurrent shift in response bias. The result is revealing because Norman's (2002) LSE was accompanied by a criterion shift. When participants change their response bias between conditions, there is always the risk that such change may affect measures of discriminability (Van Zandt, 2000). By presenting evidence of an LSE without changes in bias, we reinforce the case put forth by Norman (2002). In addition, Experiment 7 extends Norman's (2002) result by showing that LSEs are modulated by retention interval.

Contrary to our results, previous studies repeatedly found LLEs in the absence of LSEs (e.g., Murnane & Shiffrin, 1991*a*; Ratcliff et al., 1990; Ratcliff et al., 1994; Shiffrin et al., 1995). It is possible, however, that the dissociation was obtained because the interference manipulations in those studies contained features that inadvertently boosted LLEs and hindered LSEs. The manipulations may have boosted LLEs for the reasons pointed out by Dennis and Humphreys (2001). In Ratcliff et al. (1990, Exp. 6), the LLE was helped by both a proactive design at study, more detrimental to latter items in *long* lists, and output interference at test, more detrimental to long test lists. Proactive design may have also helped the LLE reported in Ratcliff et al. (1994, Exp. 3), since only the last 16 studied items in the *long* list were analysed, and the LLE in Murnane and Shiffrin (1991*a*), as attested by the smaller effect sizes found when they switched to a retroactive design.

In Shiffrin et al. (1995), where study items were semantically associated, a different factor may have boosted the LLE, namely, implicit associative responses. When study items are related to each other, it is possible that participants may implicitly generate category labels or other members of the category and encode them as if they had been studied. Implicit responses can produce LLEs because they have more opportunities to occur in *long* lists than in *short* lists, yielding more false alarms (i.e., unstudied category members are more likely to be endorsed in *long* lists). Dewhurst (2001) provided strong evidence in favour of the associative account of LLEs in categorised lists. He found that false alarms were more frequent



when the *lure* was a typical member of the category (e.g., *potato* vs. *pumpkin* for category *vegetables*) and more frequent when the category was intrinsically smaller (e.g., *days of the week* vs. *four-footed animals*). The results indicate that easily generated category members are more prone to be incorrectly recognised at test. Moreover, Dewhurst (2001) found that the frequency of *Remember* responses for *lures* increased in those conditions; the fact that participants vividly remembered non-studied exemplars suggests that they were associatively generated at study. In Shiffrin et al. (1995), false alarms to unstudied exemplars increased with category length, but hits did not change, consistent with the implicit association account. Thus, both proactive interference and implicit associative responses could have inadvertently increased the size of LLEs relative to LSEs in some previous studies.

List-strength effects, on the other hand, may have been underestimated in those studies. In Ratcliff et al. (1994), study items were presented either too fast (50 ms – 400 ms) or too slowly (2 s – 5 s). When items are presented too fast, recollection is unlikely to occur at test (Gardiner & Gregg, 1997). When items are presented too slowly, the strength manipulation is unlikely to be effective, since study times beyond 2 s produce no LSE in free recall (Malmberg & Shiffrin, 2005). This is true for studies where strength is manipulated with study time: additional study time, massed repetition or depth of processing are unlikely to yield recognition LSEs because they fail to do so in free recall (Malmberg & Shiffrin, 2005). In addition, recollection could have been reduced in some studies because *lures* were only randomly similar to *targets* (Murnane and Shiffrin, 1991a; Ratcliff et al., 1990; Ratcliff et al., 1994), allowing participants to rely on familiarity at test.

Indeed, recent studies using *unrelated lures* have replicated the early null findings. Malmberg (in press), for example, carried out a recognition study in which strong items were presented 3 times. This level of strength, which has been previously shown to elicit free-recall LSEs (Malmberg & Shiffrin, 2005; Wixted, Ghadisha, & Vera, 1997), has also been used in our Experiment 5b. Malmberg (in press) found no LSE. Likewise, Diana and Reder (2005, Exp. 2) also used only *unrelated lures* at test and found no recognition LSE, despite repeating strong items 11 times. These null effects contrast with the LSE found in Experiment 5b here, where recollection was necessary for correct performance (i.e., *SP lure* discrimination).

In the few cases where LSEs were found using only *unrelated lures*, the results may have been contaminated by floor effects on false alarms in *strong* lists (Norman, 1999, Exp. 2; 2002, Exp. 1; Experiment 7 here, SU comparison; see 4.6.4 for a discussion). By contrast, LSEs have been consistently found in item recognition (Buratto & Lamberts, 2008; Norman, 2002; Norman et al., 2008) and associative recognition (Verde & Rotello, 2004) when highly similar lures (*SP lures* and rearranged pairs, respectively) were used at test.

There are two studies, however, that used similar lures at test and yet found no LSE (Ratcliff et al., 1994, Exp. 6; Shiffrin et al., 1995). As discussed in 3.6.2, the results could have been due to low levels of recollection at test. Although unstudied exemplars from a studied category are similar to *targets* from that category, it is still possible to discriminate a *target* from a *lure* with familiarity alone. In addition, participants may refrain from using a *recall-to-reject* strategy unless explicitly told to do so (Rotello et al., 2000, Exp. 2). Finally, the ability to exhaustively search a category in order to be able to reject the *lure* as new is crucial for the effective use of a *recall-to-reject* strategy (Gallo, 2004). In other words, explicit instructions to use recollection and small category sizes are essential conditions to enable the consistent use of *recall-to-reject* in studies using list of categories (Gallo, 2004). Neither condition was satisfied in both Ratcliff et al. (1994) and Shiffrin et al. (1995). Consequently, it is possible that low levels of recollection at test masked the LSEs in those studies. As argued by Malmberg (in press), participants may be reluctant to rely on recollection – unless necessary – if they are trying to maximise *efficiency* during a recognition test. According to this view, participants routinely seek to achieve a certain degree of accuracy in the shortest amount of time.<sup>9</sup>

In brief, one can reconcile the presence of both LLEs and LSEs in Experiments 5b to 7 with previous studies where LLEs were found in the absence of LSEs by

---

<sup>9</sup> The efficiency hypothesis assumes that recollection is slower than familiarity. However, studies using the *Remember/Know* procedure found that *Remember* responses (recollection) are faster than *Know* responses (familiarity) (Dewhurst, Holmes, Brandt, & Dean, 2006; Wixted & Stretch, 2004). The discrepancy may be partially resolved by assuming that the familiarity signal elicited by a test item is available before the recollective signal (cf. Hintzman and Curran, 1994) but that participants refrain from responding based on familiarity until the attempt to recollect the item has failed. This assumption is implemented in CLS, SAC and REM (dual-process version; Malmberg, in press).

noting that those previous studies may have inadvertently boosted list-length manipulations (through proactive interference or implicit associative responses) and hindered list-strength manipulations (through the use of dissimilar lures at test).

#### 4.6.3. Implications for memory models

The present findings may help inform models of recognition memory. In the following, we assess how single-process models (BCDMEM and REM) and dual-process models (CLS and SAC) could accommodate the findings. In addition, we discuss potential implications of our findings for research on strength-based mirror effects and on the nature of recollection (i.e., whether continuous or all-or-none).

##### Single-process models

*Classic models:* Early single-process models, such as SAM, MINERVA2 and TODAM, may explain some, but not all, of the results reported here. These models correctly predict the existence of both LLE and LSE. They also predict that similar lures produce more false alarms than dissimilar lures, in accord with the data. The models, however, fail to explain the behaviour of variances across list types and the behaviour of switched-plurality false alarms upon *target* repetition. According to those models, variances in *strong* and *long* lists should be greater than in *short* lists; after all, this is how those models predict LLE and LSE (see 1.2). The results of Experiments 5a and 5b, however, showed no difference in variances across list types. Instead, LLE and LSE were caused by changes in the mean distribution values. The models also erroneously predict an increase in the number of false alarms to *SP lures* with increasing repetition of the corresponding *target* at study. The data showed either no increase or a (non-significant) decrease in *SP* false alarms for *B* items (relative to *A* items) in *strong* lists. Finally, the models predict larger LLE and LSE with increasing manipulation strength, but those effects were not observed here. In summary, the classic recognition models have problems explaining some common findings, replicated here, and some of our new findings.

*BCDMEM* — Among the modern models reviewed here, BCDMEM appears to be the one most challenged by the results. First, the model predicts no LLE and no LSE at long retention intervals. The prediction was partially supported in Experiment 7,

where no effects were found in the long interval condition, but it was not supported in Experiments 5*b* and 6, where LLE and LSE were found at both retention intervals. Second, BCDMEM predicts similar interference effects to *SP lures* and *unrelated lures* at short intervals; the results in Experiment 7, however, showed larger LLE and LSE in SSP than in SU discrimination. Third, LLE and LSE were modulated by the number of study-test blocks in Experiment 5*a* and 5*b*. BCDMEM predicts an overall decrease in performance with more blocks, due to contextual drift, but no differential effect across list type. According to the model, the difference between Experiment 5*a* and 5*b* would be reflected simply by a main effect of number of blocks; the results, however, showed an interaction, whereby LLE and LSE were larger when the number of study-test blocks was smaller.

BCDMEM could possibly fit most of our data. Dennis and Humphreys (1998) successfully modelled Murnane and Shiffrin's (1991a) results, which revealed LLEs without LSEs, by assuming that the contextual reinstatement parameter ( $d$ ) was higher in the list-length than in the list-strength condition (learning rate,  $r$ , also varied to reflect the greater strength of strong items, whereas the other parameters,  $s$  and  $p$ , remained the same across lists). Recall that  $d$  is the probability that a unit in the context vector that was active at study fails to get reinstated at test. Dennis and Humphreys (1998) assumed that  $d$  should vary with the number of unique items in a list but not with repetitions of the same item. Encoding new items would gradually change the study context such that the context reinstated at test would differ from the context present at study, harming performance. Thus, the differential loss of the original study context would account for the LLE without LSE in Murnane and Shiffrin (1991a). It would be possible to fit the LSE without LLE found in our Experiment 3 simply by reversing the assumption: repeated items harm contextual reinstatement more than new items. However, such change would be hard to justify.

A similar strategy could account for the differential effect of retention interval on *unrelated* and *SP lure* discriminability for *strong* lists. The trend for a larger LSE in the SSP than in the SU comparison at short intervals suggests that retention interval impacts differently on lures depending on their similarity to studied items. Because in BCDMEM items are represented as independent units, the model does not take into account *item similarity*. Thus, it cannot explain in terms of *target-lure* similarity

why related and unrelated lures behave differently across retention intervals. However, the model does take into account *context similarity*. The words *banana* and *bananas* or the words *banana* and *monkey* are more likely to be encoded in the same context (e.g., in the same text) than the words *banana* and *aeroplane*. As a consequence, similar items are more likely to share contexts than dissimilar items.

At test, the match between the reinstated context and the retrieved context from a *target* should co-vary with the match between reinstated and retrieved contexts from an *SP lure*. The matches should co-vary because the lure was likely to be encoded in most previous contexts where the target was encoded, even though it was not encoded in the current study context. Consequently, any factor affecting *target* matches should also affect *SP lure* matches. Increases in  $d$  (as a result of a short retention interval) should then cause both *target* and *SP lure* matches to fall; the decrease in *target* match, however, should be larger, as *targets* have more features in common with the study context than *SP lures*. Thus, higher  $d$  could account for an *SP lure* LSE. By contrast, increases in  $d$  should produce relatively little change in the match to *unrelated lures* because reinstated and retrieved contexts are only randomly similar. This can account for the *unrelated lure* LSE. The context matches in BCDMEM suggest a similar (or larger) LSE in the SU discrimination compared to the SSP discrimination because the matches to *targets* and *SP lures* decrease in tandem with a reduction in retention interval whereas the matches to *unrelated lures* remain relatively steady. This prediction, however, was not supported by our data which showed a larger LSE in the SSP comparison.

One possible solution would require the assumption that  $d$  differs between lure types, such that  $d$  would be higher for *SP lures* than for *unrelated lures*. However, there is no apparent reason why  $d$  should differ. Dennis and Humphreys (2001, p. 458) described two types of context considered critical to recognition, namely, processing and temporal context. The first is associated with the actions taken during encoding; the second is related to the passage of time. At test, both processing and temporal contexts should have changed by the same amount for *unrelated* and *SP lures*, since both lure types have not been studied (no differential processing context) and were randomly presented at test (no differential temporal context). Thus, it is not clear how BCDMEM could explain higher LSE for SSP

discrimination at short retention intervals. In sum, although BCDMEM could fit most of our data, it may need a set of implausible parameters in order to do so.

*REM* — Some of our findings may pose problems to the REM model. The fact that LSEs were consistently found in our experiments questions some of the core assumptions of the model. Recall that REM was developed, among other things, to account for null LSEs reported in previous studies (e.g., Ratcliff et al., 1990; Ratcliff et al., 1994). The model explains null LSEs through *differentiation*, the process whereby additional study of an item causes its stored trace to become more complete and, at the same time, more dissimilar to other studied items (Shiffrin et al., 1990; Shiffrin & Steyvers, 1997).<sup>10</sup> Differentiation accounts for the null LSE because strengthening some items on the study list causes the match of the other non-strengthened *targets* and the match of *unrelated lures* to decrease in tandem, resulting in no change in discriminability.

This prediction was partially borne out in Experiment 7, where no LSE was found in the SU comparison at long intervals (also the LSE observed at short intervals in the SU comparison may have been caused by floor effects; see 4.6.4). The model, however, faces a problem when trying to account for the larger and more reliable LSEs observed in SSP comparisons in Experiments 5b, 6 and 7. As with *unrelated lures*, REM predicts that *SP lures* should also decrease in tandem with their corresponding weak *targets*. That is because weak *targets* and *SP lures* are similar to each other but only randomly similar to strong items. Thus, the drop in match between a weak *target* and strong traces (i.e., differentiation) should affect weak *targets* and *SP lures* in a similar way. It is thus not clear how REM could handle both a null LSE in SU discrimination and a positive LSE in SSP discrimination.

Another finding that may challenge REM is the LLE repeatedly found in here SSP comparisons. When lures are unrelated to study items (SU comparison), REM predicts an LLE because the additional traces in memory decrease the relative impact of *target* matches in the final odds value (i.e., fewer hits) and increase the

---

<sup>10</sup> The assumption that repetitions update the representation of a single trace is in sharp contrast to the assumption made by some models, such as MINERVA2 (Hintzman, 1988) and GCM (Nosofsky, 1988), according to which repetitions result in the storage of additional copies of the item.

chances of *unrelated lure*'s matching of stored traces (i.e., more false alarms). However, when lures are highly similar to study items (SSP comparison), REM predicts no LLE. The decrease in hits with increasing list length is shadowed by a decrease in SP false alarms (see Criss & McClelland, 2006, Fig. 3, for the results of a simulation comparing hits, SP false alarms and unrelated false alarms across list lengths in an associative recognition task). Because hits and SP false alarms behave in a similar way in REM, the model predicts no net LLE in SSP comparisons.

LLEs have been found in both SU and SSP comparisons. Cary and Reder (2003, Exp. 3) used controls similar to ours and a stronger length manipulation and found a reliable LLE in an SU comparison. Moreover, Experiment 7 here yielded a weak LLE in the SU comparison, possibly masked by inter-list interference (compare Figure 4.9 with Figure 4.12). Likewise, LLEs have been found in SSP comparisons in our Experiments 5*b* and 6 and in associative recognition tasks in which lures are rearranged pairs (Criss & Shiffrin, 2004c). In sum, REM correctly predicts LLE in SU comparisons but incorrectly predicts a null LLE in SSP comparisons.

Note that the problems REM faces with *SP lures* originates from the model's asymmetry in its treatment of matches and mismatches in the likelihood ratio calculation (see Equation 1.5). Matches take into account the value of the stored features (second factor in Equation 1.5). Feature values are integers from 1 to  $\infty$  taken from a geometric distribution. Small feature values are more common in the geometric distribution and are treated as less diagnostic by the model (i.e., small feature values contribute less to the likelihood ratio). Large feature values are less common in the distribution and are treated as more diagnostic by the model (i.e., large feature values contribute more to the likelihood ratio). Thus, the match between a test item and a trace is a function not only of the number of matching features but also of the values of those features. Mismatches, on the other hand, do not take into account feature values (first factor in Equation 1.5). As a result, both common and rare features reduce the likelihood ratio by the same amount. Owing to this asymmetry, the shared features between *targets* and *SP lures* dominate over the mismatching features and, consequently, an *SP lure* behaves largely as a *target*.

To date, effects of delay have not been implemented in REM. Thus, it is difficult to evaluate whether or not the model would be able to accommodate the effects of retention interval on LLE and LSE observed here. One possibility is to include a time-sensitive parameter in the model, much like the contextual reinstatement parameter  $d$  in BCDMEM, which determines the rate at which stored feature values will be lost (i.e., a non-zero feature, which contributes to the likelihood function, becomes a zero feature, which does not contribute to the function). In this implementation, performance would fall with delay due to the loss of the original features stored at study. Consequently, interference effects would be larger at short intervals, when the interference items have more non-zero features, than at long intervals, when the interference items have fewer non-zero features. Nonetheless, a parameter that erases the values of stored features goes against the spirit of REM and previous global matching models, since one of their main tenets is that most forgetting occurs as a result of interference at retrieval rather than at storage.

REM also predicts no strength-based mirror effects to *SP lures* when strength is manipulated within and between lists. Instead, the model predicts that SP false alarms should be higher for strong items in *mixed* lists than for weak items in *mixed* lists and weak items in *pure weak* lists (Criss, 2006, Figs. 4 and 5, assuming that the similarity parameter is greater than .5). Our results, however, show that false alarms to strong-item *SP lures* either remain steady (Experiment 5a) or slightly decrease (Experiments 5b and 6) compared to weak-item *SP lures*. This incorrect prediction together with some of the issues described above have been addressed by new REM models (see next section), including a version of the model with a recall mechanism that helps to reduce the excessive likelihood ratios yielded by *SP lures*. The inclusion of recall has the potential to account for some of the results that are difficult for the original, single-process REM model to handle.

### Dual-process models

*Dual-process REM* — Although REM was conceived as a single-process model, the difficulties faced by the model in accounting for the low false alarms to highly similar lures has led to the development of a dual-process version of the model (Malmberg, Holden et al., 2004; Malmberg & Xu, 2007; Xu & Malmberg, 2007).



According to this version, a recall mechanism is triggered if familiarity ( $\Lambda$ ; see 1.4.3) exceeds a subjective criterion. If a trace is successfully retrieved, it can be compared to the test item and if they mismatch, a confident “new” response can be output. Presumably only lures that are similar to studied items will be able to invoke recall, as the familiarity produced by such lures may surpass the recall threshold.

In practice, dual-process REM has been implemented as follows: *i*) familiarity ( $\Lambda$ ) is computed as in the standard REM model; *ii*) if  $\Lambda$  does not surpass a threshold, a “new” response is produced; *iii*) if  $\Lambda$  does surpass the threshold, an attempt to recall the item is made; the probability of successfully recalling the test item is given by

$$q = ac[1 - (1 - u^*)^r] \quad (4.1)$$

where  $a$  is a scaling parameter,  $u^*$  is the probability of storing a feature,  $c$  is the probability of storing a feature *correctly* and  $r$  is the number of times an item has been studied for  $t$  seconds; *iv*) if recall succeeds (with probability  $q$ ) and the recalled trace matches the test item, an “old” response is produced; *v*) if recall succeeds and the recalled trace mismatches the test item, a “new” response is produced; *vi*) if recall fails, an “old” response is produced with high probability ( $\gamma = .9$ ).

The parameter  $a$  in Equation 4.1 varies from 0 to 1 and measures the contribution of recollection to performance: if  $a = 0$ , then  $q = 0$  and the model reverts to its single-process version; if  $a = 1$ , then  $q = c[1 - (1 - u^*)^r]$  and recollection is limited only by the quality of encoding (as  $q$  asymptotes to  $c$ ). The  $a$  parameter is relevant here because it modulates the contribution of recall at test. However, the current dual-process implementation, although inspired in the SAM model (Raaijmakers & Shiffrin, 1981), simplified several properties of its recall process. Yet SAM’s recall properties, combined with REM’s familiarity process, may be the key to our results.

In SAM, recall occurs through *sampling* and *recovery* cycles: at test, a trace is sampled; if sampling succeeds, an attempt is made to recover the trace’s contents. Once recovered, the trace can be used to accept or reject the test item. The sampling probability is proportional to the strength of the association between test item and stored trace (e.g., if the test item was studied, it will serve as a strong cue during the sampling process). The sampling probability is inversely proportional to the total

activation elicited by the test item across all stored traces. If  $C$  is the test context cue and  $I_j$  is a test item, then the probability  $P_S$  of sampling trace  $I_i$  is given by

$$P_S(I_i | C, I_j) = \frac{S(C, I_i)^{W_C} \times S(I_j, I_i)^{W_I}}{\sum_{k=1}^N S(C, I_k)^{W_C} \times S(I_j, I_k)^{W_I}} \quad (4.2)$$

where  $N$  is the number of traces in memory,  $W_C$  and  $W_I$  are attention weights and  $S(X, Y)$  are the association strengths between test cue  $X$  and memory trace  $Y$ . Note that the denominator in Equation 4.2 is the familiarity value used in the recognition version of the SAM model (see Equation 1.2). The recovery probability is given by  $P_R(I_i | C, I_j) = 1 - \exp(-S(C, I_i) - S(I_j, I_i))$ . List-length effects are predicted in this model because longer lists elicit higher activations (denominator of  $P_S$ ), resulting in smaller sampling probabilities. List-strength effects are predicted because strong items are more likely to be sampled and recovered than weak items, since their associations to the study context,  $S(C, I_i)$ , and to the other items on the list,  $S(I_j, I_i)$ , is higher and corresponds to a larger share of the total activation (denominator of  $P_S$ ).

Coupling the recall version of SAM with the standard version of REM has the potential to account for our results because, on the one hand, SAM predicts both LLEs and LSEs and, on the other hand, the role of recollection depends on the familiarity value elicited by a test item, which will be higher to similar lures than to dissimilar lures. Thus, in principle, such combined REM-SAM model could explain the larger LLE and LSE effects found in SSP comparisons than in SU comparisons. The model could also explain the between-list strength-based mirror effects of Experiments 5b and 6, insofar as recall could counteract the rise in SP lure familiarity (although the model proposed by Malmberg, Holden et al., 2004, can also explain those results). The combined model, however, may not easily account for the null effect of manipulation strength observed in our experiments, since longer and stronger lists entail greater LLEs and LSEs in the recall version of SAM.

The combined REM-SAM model could possibly explain the retention interval data through the interplay between context and delay. Gillund and Shiffrin (1984, pp. 27-30) modelled changes in study-test context by lowering the association strength  $S(C, I_j)$  between test context and stored traces, so that overall sampling and recovery probabilities decreased with changes in context, and modelled the effects of

retention interval by assuming that new items are stored during the delay, so that delay impaired recall by increasing list length. The simulations of the recall model reported by Gillund and Shiffrin (1984, Fig. 19) suggest that the probability of recall at short intervals is slightly lower than at long intervals in accord with our data. At short retention intervals, context change is large (due to participants focusing on the end-of-list context to respond) and delay is small; at long intervals, context change is either small or medium and delay is either medium or large. In most cases, the probability of recall at short retention intervals is lower than at long intervals.<sup>11</sup> This account, however, is only tentative as patterns may change with parameter settings. To summarise, the dual-process version of REM and a combined REM-SAM model may be able to explain most of our results, although firm conclusions will require actual model implementation and testing.

**SAC** — The model is able to account for some of our results. In SAC, activation elicited from a test item spreads to an episode node from several sources: the concept node representing the item, the plurality node representing the item's plurality and the context node representing the study list. If enough activation accrues to an episode node, surpassing a certain threshold, a confident "old" response is made; if activation is below threshold, a decision is made by a concept node ("old" if activation is above a concept threshold; "new" otherwise). SAC predicts LLE because adding items to a list reduces the amount of activation left in a context node to spread to episode nodes; SAC predicts LSE because strengthening some items reduces the amount of activation left in a context node to spread to the episode nodes of weak items. Thus, SAC predicts the LLEs and LSEs observed in our experiments. Moreover, because activation in SAC also depends on the recency of an item's last presentation (i.e., activation increases from baseline and quickly decreases; see 1.6.2), the model can, in principle, predict stronger effects at short retention intervals in accord with our data. Nevertheless, SAC is flexible enough to fit results showing no modulation by retention interval; the model successfully fitted Cary and Reder's (2003) data from two list-length experiments in which retention interval varied but LLEs did not change.

---

<sup>11</sup> The terms *small*, *medium* and *large* follow the levels in Gillund and Shiffrin (1984, Fig. 19, A). Recall is higher at short than at long intervals when context change is medium and delay is large.

SAC can also possibly account for the differential effects of length and strength manipulations across discrimination types (SU vs. SSP) in Experiment 7.<sup>12</sup> In SU comparisons, the familiarity of *unrelated lures* is not changed with study, since *unrelated lures* are dissimilar to study items (at least in designs such as ours, where inter-item similarity was low). So the only source of activation for *unrelated lures* at test comes from their baseline activation levels. *Unrelated lures* cannot be recalled because no episode node was created at study linking those items with the study context. Thus, the probability of responding “old” to *unrelated lures* in SAC is given by  $P(F_{unr\ lure}) = P(K)$ , where  $P(K)$  is the probability of producing a *Know* response (see 1.6.2 for a description of *Remember* and *Know* estimates in SAC).

By contrast, the familiarity of *SP lures* in SSP comparisons is indirectly changed with study because their corresponding *targets* were presented on the list. Plurality is represented separately from an item’s identity, such that each individual item is assigned to an individual concept node but all items share the same plurality node (which represents whether or not the item is in its plural form). At study, a concept node (e.g., *banana*) is encoded together with its plurality node (e.g., *s*) and its context node (e.g., *list*) to create an episode node (e.g., item *bananas* on the list). However, the link between plurality and episode sometimes is not formed at study; participants may not encode the item properly because plurality is a non-salient feature of the stimulus and performance remains poor even when participants are explicitly instructed to pay attention to it (e.g., Hintzman & Curran, 1995). An *SP lure* may be incorrectly called “old” when its episode node is activated above threshold *and* there is no link between episode node and plurality node allowing to check whether the plurality is correct (when such a link exists, both recall-to-reject and recall-to-accept strategies can be used). An *SP lure* can also be called “old” when activation in the episode node (recollection) fails to reach threshold but activation in the concept node (familiarity) surpasses its threshold. The probability of an *SP lure* is thus given by  $P(F_{SP\ lure}) = (1 - c)P(R) + [1 - P(R)]P(K)$ , where  $c$  is the probability that the plurality node is encoded together with the concept node into

---

<sup>12</sup> Firm conclusions can only be obtained with model simulation. To date, however, list-length and list-strength manipulations with *SP lures* have not been reported in SAC studies.

an episode node at study ( $c = .5$  in most SAC studies) and  $P(R)$  is the probability of a *Remember* response.

List-length and list-strength manipulations cause interference by reducing  $P(R)$ , as less activation spreads to episode nodes. *Unrelated lures* are unaffected by falls in  $P(R)$ , although  $P(F_{unr\ lure})$  may decrease due to threshold shifts in concept nodes. *SP lures*, on the other hand, are directly affected by drops in  $P(R)$ . The behaviour of *unrelated* and *SP lures* can be compared in this case;  $P(F_{SP\ lure})$  and  $P(F_{unr\ lure})$  are tied by the expression  $P(F_{SP\ lure}) = (1 - c)P(R) + [1 - P(R)]P(F_{unr\ lure})$ . The equation shows that decreases in  $P(F_{unr\ lure})$  are accompanied by smaller decreases in  $P(F_{SP\ lure})$  (i.e., the slope is less than 1 for  $0 < P(R) < 1$ ). Consequently, the differences between hits, largely determined by  $P(R)$ , and false alarms, from the expression above, are such that discriminability from *short* to *strong* lists decreases more for *SP lures* than for *unrelated lures*. In other words, SAC could potentially account for the differential interference effects across comparison types found in Experiment 7.

There is, however, one aspect of our data that could pose a problem to SAC, namely, the relatively similar LLE and LSE sizes found in Experiments 5b, 6 and 7. SAC predicts that LLEs should be either equal or larger than LSEs. This follows from the way activation spreads in the model. Adding items causes a greater drop in the amount of activation left to spread in the context node than strengthening other items. In the former, the drop in activation is linear ( $\sum_i S_{s,i}$ ), whereas in the latter it is logarithmic ( $\ln(c_L \sum_i t_i^{-d_L})$ ; see 1.6.2). It is likely that towards the end of a study list additional items are encoded less well due to factors such as loss of attention. If true, later new items should then cause less disruption in context node activation, reducing the LLE magnitude and bringing it closer to the LSE magnitude. This assumption could be implemented in SAC by allowing the strength of new concept-to-context links to vary with study position, so that later items would create weaker links. In its current form, the model can account for similar length and strength effect sizes by adopting different episode and concept thresholds across list types; changing thresholds in this post-hoc manner, however, is a less satisfactory solution.

To summarise, the SAC model is able to accommodate most of the interference effects reported here, including the differential impacts of list-length and list-strength manipulations across lure types. The model, however, have problems to account for the LLEs and LSEs of similar sizes found in Experiments 5*b*, 6 and 7.

*CLS* — Among the models reviewed here, CLS seems to have the fewest problems accounting for the results. The model predicts list-length and list-strength effects, similar effect sizes, effect modulation by lure relatedness and by retention interval. The predictions were all borne out by the data. The only results the model did not directly predict were the modulation by number of sessions (Experiments 5*a* and 5*b*) and the lack of modulation by manipulation strength (Experiments 5*b* and 6). The former may not pose a big problem for CLS, as the model may be capable of obtaining the same pattern of results by implementing a continuous memory assumption (i.e., populating the network with items prior to the experiment itself). The latter result, however, is more problematic, since CLS predicts more interference with increases in list length and strength *regardless of parameter settings*. In fact, this prediction sets CLS sharply apart from REM. Increasing item strength should boost LSE in CLS but, given enough differentiation, it could completely eliminate LSE in REM (Norman & O'Reilly, 2003, p. 638). Thus, the fact that our results showed no sign of stronger LSEs with stronger items goes against a core prediction of the CLS model. Owing to the nature of the result, however, it is important to be cautious; it is always possible that more powerful manipulations, as discussed in 4.4.4, could produce the predicted modulation.

The CLS model can account for the curvilinear ROCs produced in Experiments 5*a*, 5*b* and 6 (where only *SP lures* were tested). The curvilinear nature of the ROC in a situation that is likely to engage recollection suggests that recollection is a continuous process rather than all-or-none. Unlike in SAC, where continuous recollection is assumed, in CLS continuous recollection emerges from the model's architecture and from the level of input similarity: if study items are dissimilar, related lures are unlikely to trigger recollection and the hippocampal ROC will behave in a threshold-like manner; if, on the other hand, study items are similar, related lures are likely to trigger recollection and the hippocampal ROC becomes more and more curvilinear (compare Figs. 6B and 7 in Norman & O'Reilly, 2003,

pp. 620-621). The smooth transition of the recollection signal from threshold-like to signal-detection-like provides a framework capable of reconciling discrepant results in the literature (e.g., linear vs. curvilinear ROCs in source monitoring tasks).<sup>13</sup>

The model's prediction of LLE and LSE in the context of dissimilar study items depends crucially on the assumption that target recollection suffers interference from other list items but that lure recollection remains at floor, so that there is an overall decrease in recollection discriminability. Our stimuli were chosen so that items had little semantic similarity to each other (cosine values from Latent Semantic Analyses were all less than .4). Nevertheless, among the hundreds of words used, some similarity relations were bound to appear: 30% of the pairwise comparisons had a cosine value greater than .1; 8% had a cosine greater than .2. Thus, the average similarity in our stimulus set could map somewhere between the overlap values of 20% (threshold model) and 40% (signal detection model) reported by Norman and O'Reilly (2003). If this mapping is roughly correct, it could account for the weaker effects in SU comparisons observed here. It would be informative to carry out an experiment in which semantic relatedness (e.g., LSA cosines) is more tightly controlled in order to test the ROC predictions of the hippocampal model. Testing the model's operating characteristics, however, is complicated by the fact that threshold models may also produce curvilinear ROCs when constructed from ratings data (Malmberg, 2002). Old-new recognition tasks may thus be required.

Taken together, the results of Experiments 5 to 7 support the CLS model. However, the results are inconclusive with respect to some of CLS's predictions, such as the predicted increase in interference effects with stronger manipulations and the prediction that the recollection signal should vary from threshold-like to signal-detection-like with increasing study item similarity. These are still open questions.

One feature of the CLS model that differentiates it from all the other models discussed here is that it makes strong assumptions about the neural substrates operating during recognition. The assumptions underlying the hippocampal model are supported by neuroimaging studies that showed stronger activation of the hippocampus, relative to surrounding areas in the medial temporal lobe, when

---

<sup>13</sup> Discrepant ROC results may also arise from analyses' artifacts (Slotnick & Dodson, 2005).

participants gave high confidence “old” responses or when they responded *Remember* in associative recognition and source monitoring tasks (see Eichenbaum, Yonelinas, & Ranganath, 2007, for a review). Moreover, lesions to the hippocampus or to structures providing input to the hippocampus, such as the fornix, selectively harm recollection leaving familiarity relatively unharmed (e.g., Mayes et al., 2002; Tsivilis et al., 2008). Importantly, recent results suggest that the hippocampus can operate as a pattern-separator and mismatch-detector, being selectively activated when a test item partially matches a previously studied item but not when the test item is completely novel (Kirwan & Stark, 2007; Kumaran & Maguire, 2007a). The results support the assumptions, built into CLS’s hippocampal model, that the hippocampus carries out pattern-separation of similar input stimuli, that it can detect mismatches between recalled and current information and that recall-to-reject occurs only when there is substantial overlap between test probe and stored trace.

Some assumptions of CLS’s cortical model have also been supported. Imaging studies showed that activity in the perirhinal cortex, a structure adjacent to the hippocampus in the medial temporal lobe, is lower upon presentation of *old* items than upon presentation of *new* items in source monitoring tasks, suggesting that the perirhinal cortex may function as a novelty detector (e.g., Henson et al., 2003). Moreover, the degree of deactivation in the perirhinal cortex remained the same regardless of the success or failure of source recollection, suggesting that the perirhinal cortex may represent familiarity in the absence of recollection. Consistent with this view, Montaldi et al. (2006) found decreased activity in the perirhinal cortex with increasing levels of confidence that a picture was previously studied but no modulation of activity in the hippocampus; by contrast, activity in the hippocampus was higher only when participants reported recollecting the studied picture. Results from lesion studies strengthen this case by showing that damage to the perirhinal cortex impair familiarity but not recollection in speeded recognition and *Remember/Know* tasks (e.g., Bowles et al., 2007). Overall, these results support the idea that neocortical areas in the medial temporal lobe (e.g., perirhinal cortex) function as graded novelty detectors, sensitive to levels of mnemonic experience.

The evidence reviewed above, however, has not gone unchallenged. In particular, it has been argued that in recognition both the hippocampus and surrounding cortical



areas work together most of the time and that the difference between subjective feelings of recollection and familiarity (e.g., *Remember/Know* responses) reflect a difference between strong and weak memories rather than a difference between qualitatively distinct processes (Squire et al., 2007). Evidence for these claims come from studies showing that lesions in the hippocampus can impair both recall and recognition (Wixted & Squire, 2004), suggesting that the hippocampus may also provide useful information during familiarity-only judgements, and from studies showing that the perirhinal cortex and nearby structures previously linked to familiarity processing, such as the parahippocampal cortex and entorhinal cortex, are also active in tasks involving predominantly recollection, such as associative recognition and source monitoring (Gold et al., 2006; Kirwan & Stark, 2004). Owing to these discrepancies, it may be necessary in the future to alter some of the basic features of the CLS model, including the functional connectivity between the cortical and hippocampal models, in order to capture the apparently more fluid nature of information flow between neocortical areas (e.g., entorhinal cortex, perirhinal cortex, parahippocampal cortex) and the hippocampus.

### Strength-based mirror effects

In Experiments 5a, 5b and 6, both weak (*A*) and strong (*B*) items were tested, allowing us to assess the presence of strength-based mirror effects in our design. Strength-based mirror effects occur when hits are higher and false alarms are lower following *strong* lists (where some or all items are strengthened) compared to *weak* lists (where no items are strengthened). The relevant comparisons here involve pitting the raw measures (hits and false alarms) of *B* items in *short* lists against the measures of *B* items in *strong* lists (between-list comparison) and the measures of *A* items in *strong* lists against those of *B* items in *strong* lists (within-list comparison). For ease of reference, the relevant data from Experiments 5a, 5b and 6 for within-list comparisons are presented in Table 4.16.

Most studies that manipulated strength between lists have reported mirror effects (e.g., Hirshman, 1995; Stretch & Wixted, 1998b). Those results were also replicated here (although the false-alarm portion of the effect did not reach significance in Experiment 5a). More controversial is the status of within-list strength-based mirror

effects, particularly the conditions under which the *false-alarm portion* of the effect is produced. The issue is theoretically relevant because it is not clear why strengthening items in a study list can also affect responses to items that were *not* present on the list.<sup>14</sup> If *targets* and *lures* are somehow related (e.g., if they belong to the same semantic category), then it is possible to selectively reduce false alarms by adopting a stricter response criterion to *lures* from stronger categories. For example, if participants studied exemplars of *fruits* once and exemplars of *birds* thrice, they could improve performance on a recognition test by setting a stricter criterion to *bird* items than to *fruit* items on a trial-by-trial basis. Because several exemplars of *birds* were recently studied, it would be prudent to require more evidence (relative to *fruits*) to confidently endorse a *bird* test item as “old”.

**Table 4.16. Within-list, strength-based mirror effect (Exps. 5 / 6).**

Exp.	Weak items (A items)					Strong items (B items)				
	HR		FAR		A <sub>z</sub>	HR		FAR		A <sub>z</sub>
	M	SEM	M	SEM	M	M	SEM	M	SEM	M
<b>5a</b>	.67	.02	.32	.02	.73	.83	.02	.30	.03	.83
<b>5b</b>	.71	.03	.34	.03	.73	.85	.02	.27	.03	.83
<b>6</b>	.68	.02	.31	.03	.73	.91	.01	.29	.03	.88

*Note.* Exp. = Experiment; HR = hits; FAR = false alarms; *M* = mean; *SEM* = standard error of the mean; *A<sub>z</sub>* = sensitivity (*targets* vs. *SP lures*). Data from *strong* lists only.

Within-list criterion shifts have been observed in experiments where the nature of the task changes from trial to trial. For example, participants seem to change criteria when test lists consist of a mixture of single items and pairs of items that were differentially strengthened at study (Hockley & Niewiadomski, 2007). Participants also shift criteria when both the proportion of *new* items and the response deadlines vary at test (Heit et al., 2003). When trial-by-trial changes are less dramatic, however, participants are reluctant to change their initial criterion setting. False alarms are not lower when *lures* share the same font colour (Stretch & Wixted, 1998b) or belong to the same category (Morrell, Gaitan, & Wixted, 2002) of strong items. Moreover, false alarms do not change when the nature of the *targets* (strong vs. weak: Verde & Rotello, 2007) or the nature of the *lures* (related vs. unrelated:

<sup>14</sup> The issue of criterion setting in recognition is also important in applied settings, as attested by the current interest on decision processes in eyewitness testimony (Clark, 2003; Wells & Olson, 2003).

Benjamin & Bawa, 2004) changes conspicuously (and only once) midway through a test list. Participants also adopt the same criterion to *lures* from recently studied categories and *lures* from categories studied 40 minutes earlier (Singer & Wixted, 2006). By contrast, evidence suggests that trial-by-trial feedback (Rhodes & Jacoby, 2007; Verde & Rotello, 2007) and long delays (e.g., 2 days: Singer & Wixted, 2006) may induce participants to change their criteria dynamically during a recognition test. In sum, there is support for the view that people are reluctant to change their initial criterion, unless specific factors, such as feedback, are present at test.

Experiments 5 and 6 here included some of the factors previously shown *not* to cause within-list criterion shifts, such as a change from weak (*A*) to strong (*B*) items halfway through the test list (Benjamin & Bawa, 2004, Exp. 1; Verde & Rotello, 2007, Exp. 4) or the fact that *A* and *B* items were slightly delayed in relation to each other (Singer & Wixted, 2006, Exps. 1 and 2). Conversely, Experiments 5 and 6 included none of the factors shown to induce criterion shifts, such as feedback or long delays between item types. Thus, it is unlikely that response criteria changed between *A* and *B* items at test in our experiments. Yet, false alarms to strong items were consistently lower than false alarms to weak items in Experiments 5a, 5b and 6. In addition, the decrease in *B*-item false alarm in Experiment 5b was nearly significant ( $p = .08$ ). If response criterion did not shift between *A* and *B* items, then how can the nearly-significant mirror effect found in Experiment 5b be accounted for? One possibility is that participants were able to trigger recollection (recall-to-reject) more often when items were *B SP lures* than when they were *A SP lures*.

Evidence consistent with the idea that recall-to-reject is more frequent for strong items comes from studies showing that false alarms to *SP lures* (Light et al., 2006, Exp. 2) and to *associatively-related lures* (Benjamin, 2001, Exp. 1) is reduced as a function of repetition of the corresponding *target* items. Importantly, when recall-to-reject is disrupted by forcing participants to respond fast, false alarms to *SP lures* (Light et al., 2006, Exp. 1) and to *associatively-related lures* (Benjamin, 2001, Exp. 2) increased as a function of repetition. These findings, obtained in the context of within-list manipulations, suggest that recall-to-reject is used to counteract the increased familiarity produced by the presentation of *lures* similar to strong traces.

More direct estimates of the role of recall-to-reject in the reduction of false alarms have been obtained in studies using the *memory conjunction paradigm*. In this paradigm, participants study compound (*parent*) words (e.g., *cockpit*, *armrest*) and at test are presented with either the intact word (e.g., *cockpit*) or with a new word made from two previously studied parent words (e.g., *armpit*; conjunction lures). Of interest is the behaviour of conjunction lures when recall of parent words is facilitated. Jones (2005, Exp. 3) collected *Remember / Know* judgments for “new” responses and showed that, although false alarms did not decrease with target repetition, the proportion of *Remember* responses to conjunction lures increased, suggesting that participants used recall-to-reject more often when targets were repeated. Similarly, Odegard et al. (2005) found higher levels of recall-to-reject when conjunction lures were semantically related to its parent words (e.g., *overcoat* = *overpass* + *raincoat*) than when they were less semantically related (e.g., *payroll* = *payload* + *eggroll*); again, the increase in recollection rejection was accompanied by no overall reduction in false-alarm rates. Reduction of false alarms in the conjunction paradigm was obtained when participants were warned about the nature of conjunction lures at the time of test (Lampinen, Odegard, & Neuschatz, 2004, Exp. 1) but not when parent items were repeated (Lampinen et al., 2004, Exp. 2). In both cases, however, recall-to-reject estimates were higher. These results indicate that study repetition tends to increase recall-to-reject (which drives false-alarm rates down) and familiarity levels (which drives false-alarm rates up) to a similar degree, resulting in no net change in false alarms.

Several findings in item recognition studies with *SP lures* (Hintzman & Curran, 1995; Hintzman et al., 1992; Malmberg, Holden et al., 2004) and in associative recognition studies with *rearranged-pair lures* (Cleary, Curran, & Greene, 2001; Kelley & Wixted, 2001; Xu & Malmberg, 2007) also show little or no reduction in false alarms with target repetition. Thus, the fact that in Experiment 5*b* false alarms to *B* items nearly-significantly decreased relative to *A* items was somewhat surprising. Light et al. (2006) found a significant decrease in false alarms in their Experiment 2, where response was self-paced, but not in their Experiment 1, where response was timed (2.4 s in the long deadline condition). In our experiments, all responses were self-paced. Thus, it is possible that longer response times could allow the effects of recollection to become more apparent. Yet, this does not explain

why a similar decrease in false alarms was not found in Experiments 5*a* and 6. In particular, it is surprising that presenting a target 3 times (Experiment 5*b*) increases recollection to a greater extent than familiarity (resulting in fewer false alarms) but that the same is not observed when a target is presented 6 times (Experiment 6).

It is possible that the monotonically increasing relationship between strength and recollection becomes negatively accelerating with further repetitions, whereas the function for familiarity remains linear, as suggested by data from judgements of frequency (cf. Hintzman & Curran, 1995). This could occur if, after a few presentations, participants stop paying attention to plurality information (a distinctive stimulus feature) but continue to pay attention to the identity of the item (a prototypical stimulus feature). Consistent with this idea, when participants are forced at study to process each repeated item by adding an *s* when necessary, false-alarm rates decrease with additional repetitions (Hintzman & Curran, 1995, Exp. 4).

In sum, the results of Experiments 5*a*, 5*b* and 6 are broadly consistent with previous research showing that repetition of target items, while clearly increasing hits (see Table 4.16), has little effect on false alarms, probably due to similar increases in recollection and familiarity. Moreover, the atypical decrease in false alarms to strong lures observed in Experiment 5*b* replicates the result by Light et al. (2006, Exp. 2) and Benjamin (2001, Exp. 1), further supporting the idea that recollection can sometimes overcome familiarity in within-list manipulations of strength.

### Continuous recollection

The data from Experiments 5*a*, 5*b* and 6 was fit by unequal-variance SDT models. In those experiments, familiarity alone was not diagnostic of an item being studied. Consequently, participants would presumably have to rely often on recollection to successfully carry out the recognition task. The fact that an SDT model fitted well data from a task involving high levels of recollection indicates that the process may be better described as a continuous variable rather than all-or-none. In other words, recollection may occur with various degrees of precision rather than in a threshold fashion, where it is either triggered (with high precision) or not triggered at all.

The distinction is theoretically relevant because some dual-process models, in particular Yonelinas' (1999, 2001) model, conceive recollection as an all-or-none process, described by a two-high threshold model, and familiarity as a continuous process, described by an equal-variance SDT model. The model's recollection component predicts a linear ROC and curvilinear  $z$ ROC, whereas its familiarity component predicts a curvilinear ROC and a linear  $z$ ROC. When the linear and curvilinear ROCs are superimposed, the resulting curve is asymmetric, and when the curvilinear and linear  $z$ ROCs are superimposed the result is roughly a straight line with slope less than 1. Both asymmetric ROCs and  $z$ ROCs with slopes less than 1 are in accord with extant data (see Yonelinas, 2002, for a review).

These tenets of Yonelinas' (1999,2001) dual-process model, however, have come under increasing scrutiny due to a growing body of results showing that recognition data can be equally (or better) described by an unequal-variance SDT with no need for an additional all-or-none process (for a review, see Wixted, 2007). Moreover, tasks believed to involve recollection, such as associative recognition, source monitoring and switched-plurality item recognition produce curvilinear ROCs, contrary to the all-or-none assumption (Heathcote et al., 2006; Kelley & Wixted, 2001; Qin, Raye, Johnson, & Mitchell, 2001). Results taken as support for the threshold model, such as linear ROCs in source memory, were shown to be affected by the inclusion of trials in which participants guessed the source; ROCs become curvilinear when the guess trials are removed (Slotnick & Dodson, 2005). In addition, it has been shown that participants can attribute an item to the wrong source with high confidence (Dodson & Johnson, 1996), showing that recollections do not always produce a correct response (contrary to the all-or-none assumption). Conversely, the view that recollection can be partial is supported by studies showing that participants can vividly recollect having heard a word spoken by a male or female voice, despite failing to identify which male or female spoke the word (Dodson, Holland, & Shimamura, 1998).

The results from Experiments 5a, 5b and 6 add to this growing body of evidence by showing that people can experience degrees of recollection not only when study items are weak (presented once; A items) but also when items are strong (presented 3 or 6 times; B items). After 6 presentations, study items should elicit high levels of

recollection. Arguably, under conditions of high trace strength, recollection should behave even more strongly in an all-or-none fashion if it were in fact a threshold process. The fact that it did not, as attested by the excellent fits provided by the unequal-variance SDT model, argue against the view that recollection is all-or-none and support the view that recollection is best described as a continuous variable.<sup>15</sup>

#### 4.6.4. Limitations

The experiments reported in this thesis present design limitations that deserve to be mentioned, as they may have influenced some of the results. The first limitation concerns the number of trials per word type. In a recent review, Yonelinas and Parks (2007) suggested that at least 60 trials per word type per condition (i.e., 60 *targets* and 60 *lures*) are necessary in order to obtain well-behaved ROC curves. Regularly shaped ROCs are particularly important when trying to distinguish between models that predict curves with different shapes (e.g., threshold-based models vs. signal detection models). Although the goal of our experiments was not to determine the shape of the ROC (i.e., whether linear or concave), the fact that we used few trials per condition may have added noise to the  $A_z$  estimates, increasing the number of rejected model fits. However, the proportion of rejected models here is similar to the proportion of model rejections reported by Norman (2002, Exp. 2). In Experiment 7, data from 25% of the participants (24 out of 96) were excluded compared to 20% (16 out of 80) in Norman (2002, Exp. 2), despite the larger number of trials in the latter experiment (25 trials for *targets* and *lures* in Norman's experiment compared to 15 trials for *targets* / *SP lures* and 30 trials for *unrelated lures* in ours). Moreover, the data from rejected models was unlikely to differentially affect *long* and *strong* lists because the rejected models were evenly distributed across list types. Thus, the low number of trials here does not appear to have invalidated our results.

More likely, the scarcity of trials may have caused some of the interference effects to be underestimated. For example, the lack of an effect of list type on  $z$ ROC slopes

---

<sup>15</sup> The shape of the ROC alone cannot rule out threshold models. Malmberg (2002) showed that, when ROCs are constructed from confidence ratings, a two-high threshold model can yield both linear and curvilinear ROCs. The shape of the curve is determined by the response matrix mapping discrete internal states to overt responses. The problem with this account, however, is complexity: the number of parameters necessary to fit a curvilinear ROC with a threshold model tends to be larger than the number of parameters needed to fit the curve with an unequal-variance SDT model.

(Experiments 5a, 5b) could be partly due to the unreliability of the estimates: Macmillan et al. (2004) showed that slope estimates were extremely variable and recommended at least 200 trials per word type in order to achieve acceptable levels of accuracy and precision. Clearly, more trials per condition would reduce noise and increase the reliability of the data reported here.

The second limitation of our experiments was the excessive number of study-test blocks per session. Participants underwent three blocks per session (one for *short* lists, one for *long* lists and one for *strong* lists), with list type counterbalanced across participants (i.e., 6 different block orders). Multi-list sessions were chosen for practical reasons; participants would otherwise have to attend 6 sessions, one for each list type and retention interval combinations. Although block order did not interact with list type in most experiments, in Experiment 7 it nearly did so [ $F(10,132) = 1.76$ ,  $MSE = 0.01$ ,  $p = .08$ ]; the interaction suggested that sensitivity for a list type tended to be better when that list type was in the first study-test block. This sort of inter-list interference has been previously observed. Diana and Reder (2005, Exp. 1) had to change their list-strength design from multi-list to single-list after having tested 39 participants because performance in the second block was much poorer than in the first block regardless of list type. Likewise, Norman (1999, Exp. 2) found that his recollection LSE, measured with the *Remember/Know* procedure, was lower in the second half than in the first half of his four-block experiment. Taken together, these results suggest that inter-list interference may reduce or even eliminate any signs of list-strength interference. Inter-list interference may do so by reducing the impact of strong items on weak items in *strong* lists, by reducing performance in the baseline condition (*short* lists), or both.

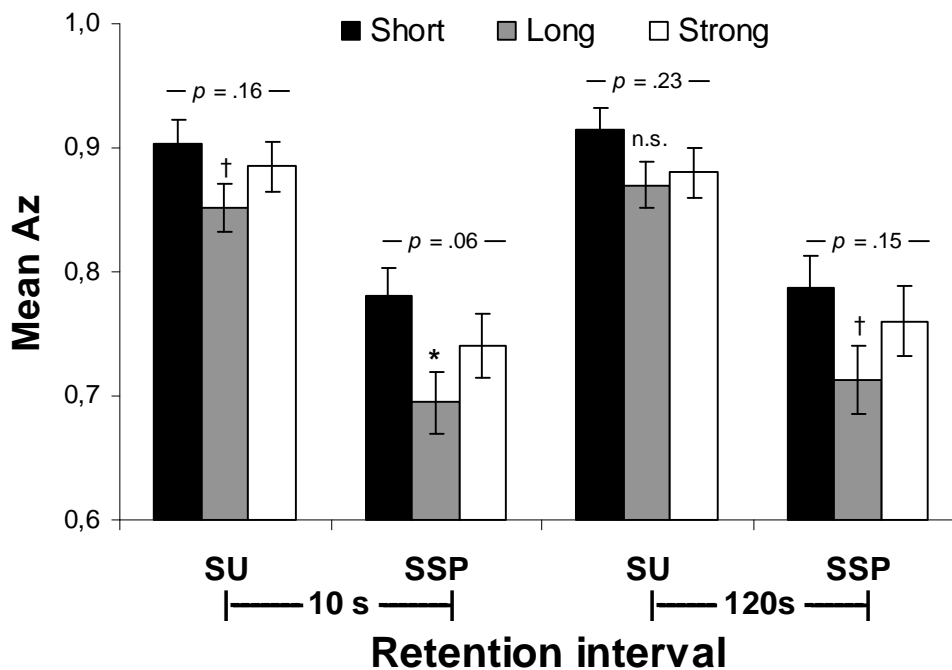
Inter-list interference has also been observed in list-length studies. Gronlund and Elam (1994) found larger LLEs in a single-list compared to a multi-list study. Similarly, we found in Experiment 7 that the null LLE observed in the SSP comparison, with data collapsed across study-test blocks, was significant with data from the first block only.<sup>16</sup> The fact that the LLEs in Experiment 7 tended to be slightly smaller than the LSEs with multi-block data but slightly larger with first-

---

<sup>16</sup> The null LLE in Experiment 3, however, cannot be attributed to inter-list interference because the results from the first study-test block were very similar to results from all three blocks (see 3.4.2).



block data suggests that list-length manipulations might be more sensitivity to the effects of multiple lists than list-strength manipulations. Figure 4.12 shows these results. It is important to note that a considerable amount of power was lost in the first-block analysis, as the within-participant comparison across lists ( $N = 72$ ) had to be converted into a between-participant comparison ( $N_{short} = 26$ ,  $N_{long} = 24$ ,  $N_{strong} = 22$ ). Consequently, previously significant LSEs, are not significant here. Crucially, only in Experiment 7 have the results changed noticeably between multi-block and first-block analyses. For Experiments 1 to 6, the patterns remained unaltered. Yet the fact that results did change in Experiment 7 indicates that strong claims about the relative magnitudes of LLEs and LSEs cannot be made with the available data.



**Figure 4.12. Sensitivity data from first study-test block (Exp. 7).**

Sensitivity for *long* lists was lower than for *short* lists in SSP comparisons across retention intervals but only marginally lower in SU comparisons. Sensitivity for *strong* lists was (non-significantly) lower than for *short* lists in both SU and SSP comparisons, but numerically larger in SSP comparisons and at short retention intervals. Significance values ( $†$ ,  $*$ ) refer to performance relative to *short* lists; the  $p$ -values at the top represents the significance of the main effect in the one-way ANOVAs across list types. SU = studied vs. unrelated lures; SSP = studied vs. switched-plurality.  $A_z$  = sensitivity (area under ROC). Error bars = SEM. n.s. non-significant;  $† p < .10$ ;  $* p < .05$ .  $N = 72$ .

A final limitation in our experiments refers to floor effects. Because list-strength manipulations almost invariably lead to drops in both hits and false alarms, it is important to assess whether false alarms in *strong* lists are not reaching floor more often than in *short* and *long* lists. If they do, then the net result may be a spurious LSE: hits fall more than false alarms, yielding lower  $d'$  values. Note that floor

effects are less of a problem in list-length studies because LLEs are usually characterised by higher false alarms in *long* lists. One can measure floor effects by counting the number of data entries where  $F = 0$ ; zero false alarms are usually corrected to avoid infinite  $d'$  values.<sup>17</sup>

Although corrections occurred to both *targets* (when  $H = 1$ ) and *lures* in all experiments, we believe that floor effects might have significantly influenced only one result. The LSE found in the SU comparison (10 s) in Experiment 7 may have been inflated by a floor effect because the number of false-alarm corrections to *unrelated lures* was much higher for *strong* lists than for *short* and *long* lists. Corrections to *targets* and *SP lures*, by contrast, affected less than 5% of the data. Norman (2002, Exp. 1) reported a similar problem when analysing  $d'$  data from *Remember* responses. Because *Remember* false alarms were rare in his experiment, thereby requiring floor corrections, and because there were more zero false alarms in *strong* lists than in *short* lists, it is possible that part of his *Remember* LSE could have been overestimated. Owing to the role of floor effects in the SU comparison in Experiment 7, we think it is safer to dismiss the LSE in that condition as artifactual.

In summary, the experiments reported in this thesis present some important limitations in terms of design. These limitations, however, do not undermine most of our results. It is important to address those issues in future work by increasing the number of test trials per word type, by carrying out fewer study-test blocks per session (preferably only one) and by making the recognition task slightly more difficult to reduce the number of false-alarm corrections. In addition, modern statistical techniques, such as bootstrapping (Schooler & Shiffrin, 2005) and Bayesian analysis (Dennis, Lee, & Kinnell, submitted) could be used to reanalyse the present data. Bootstrapping, in particular, seems promising because it permits tackling the thorny issue of sparse data, which was ubiquitous in our experiments.

---

<sup>17</sup>  $d'$  becomes infinite when  $F = 0$  because  $d' = z(H) - z(F)$  and  $z(0) = -\infty$  (see also 2.3.2).

## Chapter 5. General Discussion

When I learn the name of a student, I forget the name of a fish. **David Starr Jordan (1851-1931)** \*

### 5.1. Summary

Forgetting is an important property of the human memory system. Explaining why people forget may shed light on both basic questions (e.g., how memories are lost) and applied questions (e.g., how to improve performance in the classroom or in the workplace). Forgetting can occur at encoding (e.g., new traces damaging old traces) or retrieval (e.g., traces competing to reach awareness). In this thesis, we investigated how adding new items to memory affects memory for the other items already stored and how strengthening some items through repeated exposure affects memory for the non-strengthened items. Under several conditions, these list-length and list-strength manipulations were shown to affect performance, such that recognition of any item in a long list or any weak item in a mixed list was worse than recognition in a control list. In the following, we summarise our findings and their implications for memory models.

#### 5.1.1. Empirical implications

Experiments 1 to 4 followed the design in Dennis and Humphreys (2001) and Experiments 5 to 7 followed the design in Norman (2002). Those studies were chosen because they produced results that contradicted a wealth of previous research. We set out to investigate why they yielded those results, trying to identify some of the boundary conditions behind list-length and list-strength effects. The experiments here differed from Dennis and Humphreys' (2001) because we varied target-lure similarity. The experiments also differed from Norman's (2002) because we varied list length. Finally, the experiments differed from both studies because we varied encoding task (size vs. pleasantness), retention interval (short and long) and manipulation strength (strong items shown 3 times vs. 6 times; long lists 2 times longer than short lists vs. 3.5 longer).

---

\* Professor of Ichthyology and president of Stanford University (cited in Anderson et al., 2000).

First, the results showed that the null LLE and LSE reported by Dennis and Humphreys (2001) were probably caused by low levels of recollection at test. In Experiments 1 to 4, no significant LLEs and LSEs were found when lures were unrelated to targets but LLE (Experiment 4) and LSE (Experiments 3 and 4) were found when lures were highly related to targets.<sup>1</sup> Second, the results of Experiments 1 to 4 indicated that the null effects in Dennis and Humphreys (2001) were probably not a consequence of participants' use of covert rehearsal strategies at study, since null effects were also found in Experiments 1 and 2 here where participants were encouraged to move from item to item as quickly as possible (self-paced encoding task). Third, the LSE in Norman (2002) was probably not boosted by the use of a longer list in the strong condition, as LSEs of similar magnitude were found here even after adding a tightly matched list-length manipulation (Experiments 3 and 7; SSP comparison).

Fourth, encoding task was probably not the critical factor behind the discrepant results of Dennis and Humphreys (2001) and Norman (2002): although there was a main effect of encoding task in Experiment 1, suggesting that size judgements caused more interference than pleasantness judgements, there was no interaction with list type. Fifth, inter-list interference may mask length and strength effects. When participants underwent 6 study-test blocks in one session, no interference effects were found (Experiment 5a). By contrast, when they underwent 3 blocks per session on different days, both LLE and LSE were found (Experiment 5b). Analyses of signal-detection parameters showed that there was no difference in variance across list types; the interference effects were instead accounted for by changes in estimated target means.

Sixth, retention interval can modulate LLE and LSE magnitudes. In Experiment 2, retention interval was fixed at 180 s for long and strong lists and no effects were found; in Experiment 3, there was no retention interval (0 s) and an LSE was found. Similarly, in Experiment 7, LLE and LSE were larger when retention

---

<sup>1</sup> In Experiment 7, the null LLE in the SU comparison was probably due to inter-list interference, since analyses of first-block data showed a significant LLE. The LSE observed in the SU comparison, on the other hand, was probably artifactual due to floor effect on false alarms.

interval was short (10 s) compared to when it was long (120 s). However, not all experiments where interval was manipulated produced the effects (Experiments 4, 5 and 6). The unreliability of retention interval as a modulator of interference is somewhat surprising given that most models predict stronger effects when interfering items are stronger, which should happen when those items were studied more recently (i.e., short interval condition). Seventh, increasing the length of the long list or the repetitions in the strong list had little effect on the magnitudes of LLE and LSE (Experiments 3 vs. 4 and Experiments 6 vs. 5*b*). Note, however, that the increase in list length from Experiment 3 to 4 produced an increase in LLE. The small effect of manipulation strength was unexpected given the prediction that stronger effects should follow stronger manipulations.

Overall, our results show that LLE and LSE in recognition are robust effects that do not depend heavily on design features such as encoding task or encoding time but that depend on the relative contribution of recollection at test and can depend on the size of the experimental session and on the length of the retention interval.

### 5.1.2. Theoretical implications

Some of our experiments may have consequences for current memory models. We focused on BCDMEM, CLS, REM and SAC, because they have built on the insights of classic *global memory models* and because they are process models (i.e., they implement mechanisms for encoding, storage and retrieval of traces).

First, the fact that interference was larger in conditions where recollection was more important than familiarity constitutes evidence in favour of CLS and SAC, which predict a higher impact of interference manipulations on their recollection components. The result provides evidence against BCDMEM, which does not differentiate recollection-based from familiarity-based decisions. The fact that LSEs were repeatedly found in our experiments also provides evidence against REM, which predicts little or no LSE due to its *differentiation* mechanism. Second, the finding that retention interval can modulate interference effects supports BCDMEM and CLS (and possibly SAC). However, the differential

effects of retention interval across lure types argue against BCDMEM, which predicts similar effects of interval, regardless of lure similarity.

Third, the null impact of stronger manipulations on the magnitudes of LLE and LSE argues against CLS and SAC, which predict more interference with longer and stronger lists, and against REM, which predicts increasing LLEs with longer lists but no LSEs with stronger lists. In contrast, BCDMEM predicts no effects at all, regardless of manipulations strength. Because these are null results, however, they do not provide strong evidence against CLS, SAC and REM.

Finally, length and strength manipulations caused similar impairments on memory performance (except in Experiment 3), providing support for CLS, which predicts similar changes in its network connection weights for extra items and repeated items. The result is not consistent with SAC, which predicts larger length than strength effects, as a consequence of the way activation spreads in the model, and REM, which predicts no LLE in SSP comparisons.

Taken together, our results support CLS the most and BCDMEM the least. CLS is the only of these four models that poses that interference occurs at encoding rather than at retrieval. The fact that CLS accounts well for most of our data suggests that interference-at-study models are viable and that competition between traces at encoding, in addition to competition at retrieval, can partly account for forgetting in memory tasks. By contrast, BCDMEM is the only of the four models that poses that interference is due solely to competition between the contexts in which an item has been previously seen. The fact that BCDMEM cannot explain most of our data indicates that context interference is not the only factor contributing to forgetting. Models that assume forgetting as a result of competition between *items*, either at study or retrieval, do a better job at explaining the results in this thesis than models assuming that forgetting is all due to competition between *contexts*.

## 5.2. The role of retrieval practice

There is a growing body of evidence suggesting that the mere act of retrieval can cause the forgetting of related material (for a review, see Anderson, 2003). This phenomenon, named *retrieval-induced forgetting* (RIF), has been observed with a variety of stimulus classes (e.g., verbal: Anderson, Bjork, & Bjork, 1994; visuo-spatial: Ciranni & Shimamura, 1999), memory paradigms (e.g., cued recall: Anderson, Bjork, & Bjork, 2000; item recognition: Hicks & Starns, 2004; free recall: Macrae & Macleod, 1999; associative recognition: Verde, 2004) and experimental settings (e.g., fact retrieval: Anderson & Bell, 2001; eyewitness testimony: Shaw, Bjork, & Handal, 1995).

RIF has commonly been found with the *retrieval practice paradigm* (Anderson et al., 1994). In this paradigm, participants study items from lists of category exemplars (e.g., *fruits: banana, apple; professions: dentist, plumber*) and subsequently practice items from some categories (e.g., *fruits: banana*) but not from others (e.g., *professions*). Crucially, some items from the practised categories are not practised (e.g., for *fruits*, *banana* is practised but *apple* is not). RIF is said to occur when the final recall of unpractised items from practised categories (e.g., *fruits: apple*) is worse than recall of unpractised items from unpractised categories (e.g., *professions: dentist*), despite the fact that those words were presented the same number of times at study.

The dominant theoretical view posits that RIF is caused by an inhibitory process that suppresses memories of competitors at retrieval time, thus facilitating the recovery of target information (Anderson, 2003). Thus, during the retrieval practice of *fruit – banana*, strong category associates, such as *apple*, may come to mind and need to be suppressed to allow the successful recall of *banana*. This suppression shows in the final memory test, where memory for non-practised category associates is lower compared to memory for non-practised items from categories where such competition at retrieval has not occurred. Evidence for the inhibitory account comes from studies showing that RIF is a cue-independent process (Anderson & Spellman, 1995; but see Perfect et al., 2004, for data suggesting that RIF may be cue-dependent). That is, it is the exemplar itself that

is affected during retrieval practice, not the association between category and exemplar (the trace for *banana* is affected, not the association *fruit – banana*). Cue-independence is relevant for inhibitory models because most non-inhibitory models (e.g., Mensink & Raaijmakers, 1988) assume that forgetting is caused by strength-dependent competition at retrieval, a cue-dependent process. Subsequent studies revealed boundary conditions on RIF: forgetting disappears when the interval between practice and test is longer than 24 hours (MacLeod & Macrae, 2001), when participants are instructed to inter-relate the exemplars on the list (Anderson & McCulloch, 1999) and when strengthening in the practice phase is achieved by item repetition rather than recall (Anderson, Bjork et al., 2000).

The strength manipulation used in Experiments 1 to 7 renders our procedure in some respects similar to the retrieval-practice paradigm. By studying interference items repeatedly in the study phase, participants effectively carried out retrieval practice. Each subsequent presentation of an interference item may have triggered the retrieval of the fact that the item was previously studied (i.e., a form of recursive reminding: Hintzman, 2004). As a result, it could be argued that the LSE found here was caused by retrieval-induced forgetting. In other words, participants were worse at remembering weak items in strong lists compared to weak items in short lists because the retrieval process triggered by repeated presentations in the former may have suppressed memory for the other items. The RIF account contrasts with the accounts from the CLS (item interference) and BCDMEM (context interference) models, where interference is assumed to occur at storage, and with the account provided by SAC, where interference is assumed to be a cue-dependent process.

Although we cannot entirely rule out the possibility that our results were caused by retrieval-induced forgetting, it is unlikely that RIF played a major role here. Anderson et al. (1994) showed that RIF is strongly dependent on the strength of competitors activated at retrieval. In the retrieval-practice paradigm, strong competition has usually been achieved through the use of lists of highly related words, such as multiple instances from the same semantic category. However, the tasks used here involved lists of unrelated words. Because our items were unrelated, competition at retrieval was probably weak, reducing the impact of



retrieval-induced processes. Moreover, the modulation of LSE by retention interval observed in Experiments 2 and 3 and in Experiment 7 is generally not observed in RIF studies. RIF tends to change little with retention intervals as long as 20 minutes, whereas LSE was reduced here when retention interval varied from 0 s to 180 s. Thus, although RIF may have contributed to some of our list-strength findings, it cannot account for all the results.

### 5.3. Further Directions

The list-length and list-strength effects described in this thesis tended to be more pronounced in conditions where recollection was likely to operate. Similar requirements apply to associative recognition (intact vs. rearranged pairs), cued recall and source monitoring. In all those paradigms, familiarity of an *old* item is not sufficient to allow successful discrimination from a *new* item, and a process akin to recall seems necessary. Given the similarities between the tasks used here and the aforementioned paradigms, one possibility is that LLEs and LSEs in those paradigms could also be modulated by retention interval. In the following, we discuss this and other possibilities in cued recall and associative recognition.

#### 5.3.1. Cued recall

List-strength effects were first found in free recall tasks (Tulving & Hastie, 1972) and were subsequently replicated in several studies (Malmberg & Shiffrin, 2005; Ratcliff et al., 1990; Wixted et al., 1997). However, the cause of the effect is still a matter of debate, as is the empirical status of LSE in cued recall. Establishing the presence of an LSE in cued recall is important because the argument for an LSE in recognition relies on the assumption that recognition is mediated by a recollection process that is akin to cued recall (i.e., participants use the test item as a cue to recall the plurality of the item). According to this view, a recollection-dependent LSE in recognition should be observed if and only if a cued-recall LSE can also be observed. A failure to find an LSE in cued recall would thus challenge the assumption that recollection is a recall-like process. Here we review some conflicting results in free and cued recall and sketch an experiment to test whether list-strength effects in cued recall really exist.

### Output interference and LSE

Models that assume interference at retrieval, such as SAM, can explain LSEs in free recall and cued recall by a mechanism in which strong items tend to be sampled and recovered at test more often than weak items because strong items are more strongly associated with the study context. Consequently, weak items in *mixed* lists are recalled less often than weak items in *pure weak* lists, indicating an LSE. Thus, SAM predicts positive list-strength effects in recall.

In free recall, participants are free to recall both strong and weak items in any order. In cued recall, participants have to recall a target according to the sequence of cues presented at test. The fact that recall occurs in any order in free recall but not in cued recall has consequences for the list-strength effect. The probability of recalling an item along a testing sequence decreases for later positions in the sequence (output interference: Roediger & Schmidt, 1980). In *mixed-strength* lists, strong items tend to be recalled before weak items. As a result, weak items suffer more interference than strong items in *mixed* lists because they are recalled later on the list. Comparing the percentage of weak items recalled in *pure weak* lists with the percentage of weak items recalled in *mixed* lists reveals lower recall in the latter, characterising an LSE in free recall. Thus, output interference could account for the LSE in free recall. In contrast, output interference is presumably less effective in cued-recall tasks because the order of the cues at test is usually random, so that cues to strong and weak items end up being evenly spread along the testing sequence. As a result, smaller LSEs should be found.

Consistent with the output-interference hypothesis, Ratcliff et al. (1990, Exp. 6) found that free-recall LSEs tended to be larger than cued-recall LSEs. Ratcliff et al. (1990) used as a measure of LSE the ratio of the performance measure for strong to weak items in *mixed* lists divided by the same ratio in *pure* lists (ROR; ratio of ratios); an LSE is said to occur if ROR is significantly greater than 1. Ratcliff et al. (1990, Exp. 6) found an ROR of 1.24 for cued recall and 1.62 for free recall, supporting the idea that output interference may contribute to list-strength effects in recall tasks.

A more direct test of the output-interference hypothesis was conducted by Bauml (1997), who found that free-recall LSEs can be eliminated if output order is controlled. Output order can be controlled in free-recall tests by using a cue that uniquely identifies a *target* (e.g., the item's semantic category and first letter). Bauml (1997) found that an LSE was present when strong items were tested first in *mixed* lists but not when weak items were tested first. He concluded that free-recall LSEs are caused by interference occurring at retrieval through a process of *suppression*, whereby competitors are inhibited in order to facilitate the recall of a target item. According to the suppression view, successful recall entails the temporary inhibition of competing items. Bauml (1997) argued that if LSEs in recall were caused by *strength-dependent competition*, whereby the increase in the strength of association between an item and the study context entails the decrease in the association of the remaining items to the study context, then one would expect little or no impact of events taking place at retrieval, such as output interference, on the magnitude of the LSE. Thus, retrieval-dependent suppression rather than item-to-context competition, could account for the LSE in free recall.

Bauml (1997) made a similar case for cued recall. He noted that the LSEs in Ratcliff et al. (1990) were driven by higher recall of strong items on *mixed* lists than on *pure strong* lists. The other side of the effect – more recall of weak items in *pure weak* lists than in *mixed* lists – was not significant. Bauml (1997) argued that such pattern would be expected if one assumes that output interference depends on successful recall, and that successful recall is more likely in *pure strong* lists than in *mixed* lists. In sum, Bauml's (1997) results suggest that both free- and cued-recall LSEs can be explained in terms of suppression events taking place at retrieval, rather than as a consequence of strength-dependent competition which, despite being enacted at retrieval, is the result of events taking place at study (i.e., differential associations of items to the study context).

SAM can still account for Bauml's (1997) data because it implements the idea that items can also be learned at test. According to this process, items successfully recalled at test become more strongly associated to the test context and, therefore, more likely to be sampled and recovered in subsequent recall

attempts (Raaijmakers & Shiffrin, 1981, p. 110). This mechanism mimics output interference. Thus, SAM takes into account output interference as a contributing factor towards LSEs in recall. Bauml's (1997) results, however, suggest that when such learning-at-test mechanism is factored out, strength-dependent competition should cause minimal interference in recall. Thus, Bauml's (1997) data still poses a challenge to models that assume that strength interference occurs as a result of competition for context.

Bauml's (1997) data also poses a challenge to models that assume interference at study. In CLS, for example, interference occurs at study because new items or strong items take up connection weights in the network, disrupting previously stored connections. Note that in most models reviewed here, previously encoded traces are left intact after the strengthening of another trace; interference occurs due to sampling competition or other non-deleterious processes. In models like CLS, however, interference is deleterious. If list-strength effects in free and cue recall can be eliminated at will when output interference is controlled, then it stands to reason that weak items were not affected by other items in a deleterious manner during study.

Because Bauml's (1997) data poses challenges to both interference-at-study (e.g., CLS) and interference-at-test models (e.g., SAC, REM), it is important to take a closer look at the procedure used in his study. In the following, we argue that Bauml's (1997) experiments may have underestimated strength-dependent competition by strengthening items through study time rather than item repetition. In addition, we propose a test of the suppression account.

### Is there LSE in cued recall?

In Bauml's (1997) study, strength was manipulated by increasing study time from 2 s (weak items) to 6 s (strong items). Previous studies, however, have shown that manipulating strength with study time leads to lower interference effects than manipulating strength with study repetition, despite sometimes similar levels of strengthening (see Malmberg & Shiffrin, 2005, for a discussion).

For example, the cued-recall LSEs in Ratcliff et al. (1990) were not significant in Experiment 3, where strength was manipulated with study time ( $ROR = 1.07$ ), but were significant in their Experiment 4, where strength was manipulated with item repetition (strong items were presented 4 times; all  $RORs > 1.24$ ).

In the few cases where study time did produce slightly larger effects than repetition, it is difficult to draw conclusions due to possible confounds. For instance, Kahana, Rizzuto and Schneider (2005) found a cued-recall LSE in their Experiment 1, where study time was manipulated ( $ROR = 1.42$ ), but a smaller LSE in Experiment 2, where item repetition was manipulated ( $ROR = 1.34$ ). Experiments 1 and 2, however, differed in that list type was manipulated within participants in the former and between participants in the latter, possibly reducing the power of the repetition manipulation. Moreover, in both experiments, study-test lag was not controlled across list types such that the average time between study and test of a weak item was shorter in *pure weak* lists than in *mixed* lists, possibly inflating effect sizes.

More concrete evidence of the differential effects of time and repetition comes from Malmberg and Shiffrin (2005) who found LSEs in free recall when strong items were presented 3 times but not when study time increased from 1 s to 3 s. Malmberg and Shiffrin (2005) interpreted their results as evidence that the amount of context encoded at study, which is essential to LSE in models such as SAM, increases with spaced study, but not with massed study.

Bauml (1997) used massed study to implement item strengthening and showed only a small LSE in the output-interference condition. Moreover, the effect was significant only for the comparison between strong items across *pure* and *mixed* lists. Because Bauml's (1997) goal was to demonstrate a null effect, it is crucial that a reliable effect is obtained first with the factor deemed artifactual (e.g., output order) before demonstrating the null effect of interest; otherwise the lack of effect may be attributed to factors other than the one of interest.

In sum, given that the evidence for the output-interference account of LSE is based on suboptimal manipulations of strength and given that previous LSEs in

cued recall may have been affected by confounds, we believe that it is reasonable to conclude that LSE in cued recall still remains an open question. Establishing whether the effect in recall is due to strength-dependent competition or retrieval-dependent suppression is theoretically important because numerous models adopt strength-dependent competition as the mechanism behind interference effects. An LSE in cued recall would also support models that pose interference at study.

LSE could be investigated in the future by adopting the design features used in Norman (2002) and in here, including the spaced repetition of strong items.<sup>2</sup> In addition, it would be relevant to assess whether cued-recall LSEs behave more like retrieval-dependent phenomena, such as RIF (see 5.2), or more like strength-dependent phenomena, like LLE in recognition. Manipulating retention interval may provide a useful testing ground: a result showing modulatory effects of retention interval on cued-recall LSEs would constitute evidence for the strength-dependent competition hypothesis and against retrieval-dependent suppression hypothesis, since it is known that retrieval-induced forgetting is long-lasting.

### 5.3.2. Associative recognition

List-strength effects in associative recognition were first reported by Verde and Rotello (2004). In their experiment, sets of overlapping pairs (e.g., AB, AC, DB) were presented at study. Some sets had pairs presented only once (control items); other sets had half of their pairs presented once (weak pairs) and half presented 3 times (strong items). The same words appeared in both weak and strong pairs, so that the difference between pairs was not in their individual words but instead in the strength of their *associations*. Intact and rearranged pairs were presented at test. The results showed lower discriminability for weak pairs compared to control pairs, constituting a list-strength effect.

Verde and Rotello (2004, Footnote 2) noted that the strength of their interference manipulation may have helped to obtain reliable LSEs in associative recognition

---

<sup>2</sup> Norman (2002) pointed out that previous cued-recall studies did not have encoding tasks. The absence of encoding tasks – such as the size judgement task, designed to increase trace overlap – may have reduced interference in those studies even further. In RIF studies, encoding tasks can in fact change retrieval-related suppression into facilitation (Anderson, Green, & McCulloch, 2000).

compared to previous results in cued recall, where unreliable effects were found. Indeed, not only *lures* were highly similar to *targets* (i.e., rearranged pairs) but also *targets* were similar amongst themselves (i.e., overlapping sets of pairs). According to the CLS model, LSEs should be higher when both *target-lure* and *target-target* similarity are high compared to when only *target-lure* similarity is high, as in the experiments here (Norman & O'Reilly, 2003, Figs. 9 and 19).

The prediction of the CLS model together with our results in Experiments 3 and 5*b* suggest that an LSE would also be found in associative recognition even if non-overlapping pairs were used at study and even if new pairs were used at test. Thus, one possible way forward would be to systematically vary, in the same experiment, both *target-lure* and *target-target* similarity and observe whether progressively larger LSEs would be found from low *target-target* / low *target-lure* similarity (low overall interference) to high *target-target* / high *target-lure* similarity (high overall interference). CLS makes the unique prediction that not only LSE should be found in SSP comparisons (intact vs. rearranged pairs) when *target-target* similarity is either low or high but also that LSE should be found in SU comparisons (intact vs. new pairs) when *target-target* similarity is high. Finally, retention interval could also be manipulated to test the general prediction that interference effects should be higher when strong items were more recently studied (i.e., short retention interval condition).

To summarise, the uncertain status of LSE in cued recall and the possible modulatory role of similarity in associative recognition invite further research. Future experiments, such as the ones sketched above, would allow testing some of CLS's predictions, possibly fostering additional model development.

## References

- Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences*, 22, 425-443.
- Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, 49, 415-445.
- Anderson, M. C., & Bell, T. (2001). Forgetting our facts: The role of inhibitory processes in the loss of propositional knowledge. *Journal of Experimental Psychology General*, 130, 544-568.
- Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin and Review*, 7, 522-530.
- Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term-memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20, 1063-1087.
- Anderson, M. C., Green, C., & McCulloch, K. C. (2000). Similarity and inhibition in long-term memory: Evidence for a two-factor theory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 26, 1141-1159.
- Anderson, M. C., & McCulloch, K. C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 608-629.
- Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review*, 102, 68.
- Arndt, J., & Reder, L. M. (2002). Word frequency and receiver operating characteristic curves in recognition memory: Evidence for a dual-process interpretation. *Journal of Experimental Psychology: Learning Memory and Cognition*, 28, 830-842.
- Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. . In D. H. Krantz, R. C. Atkinson, R. D. Luce & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 1, pp. 242-293). San Francisco: W. H. Freeman and Co.



- Barnes, M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology*, 58, 97-105.
- Bauml, K. H. (1997). The list-strength effect: Strength-dependent competition or suppression? *Psychonomic Bulletin and Review*, 4, 260-264.
- Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 941-947.
- Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, 51, 159-172.
- Bliss, T. V. P., & Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361, 31-39.
- Boldini, A., Russo, R., & Avons, S. E. (2004). One process is not enough! A speed-accuracy tradeoff study of recognition memory. *Psychonomic Bulletin and Review*, 11, 353-361.
- Bowles, B., Crupi, C., Mirsattari, S. M., Pigott, S. E., Parrent, A. G., Pruessner, J. C., et al. (2007). Impaired familiarity with preserved recollection after anterior temporal-lobe resection that spares the hippocampus. *Proceedings- National Academy of Sciences USA*, 104, 16382.
- Brockdorff, N., & Lamberts, K. (2000). A feature-sampling account of the time course of old-new recognition judgments. *Journal of Experimental Psychology: Learning Memory and Cognition*, 26, 77-102.
- Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, 2, 51-61.
- Brown, M. W., & Xiang, J. Z. (1998). Recognition memory: Neuronal substrates of the judgement of prior occurrence. *Progress in Neurobiology*, 55, 149-189.
- Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods Instruments and Computers*, 35, 11-21.
- Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science*, 18, 40-45.
- Buratto, L., & Lamberts, K. (2008). List strength effect without list-length effect in recognition memory. *Quarterly Journal of Experimental Psychology*, 61, 218-226.

- Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, 49, 231-248.
- Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 1403-1414.
- Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition*, 20, 231-243.
- Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology*, 17, 629-654.
- Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and Review*, 3, 37-60.
- Clark, S. E., Hori, A., & Callan, D. E. (1993). Forced-choice associative recognition: Implications for global-memory models. *Journal of Experimental Psychology: Learning Memory and Cognition*, 19, 871-881.
- Cleary, A. M., Curran, T., & Greene, R. L. (2001). Memory for detail in item versus associative recognition. *Memory and Cognition*, 29, 413-423.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33, 497-505.
- Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language*, 55, 461-478.
- Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language*, 55, 447-460.
- Criss, A. H., & Shiffrin, R. M. (2004a). Context noise and item noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review*, 111, 800-807.

- Criss, A. H., & Shiffrin, R. M. (2004b). Interactions between study task, study time, and the low-frequency hit rate advantage in recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30, 778-786.
- Criss, A. H., & Shiffrin, R. M. (2004c). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition*, 32, 1284-1297.
- Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory and Cognition*, 28, 923-938.
- Curran, T., & Cleary, A. M. (2003). Using ERPs to dissociate recollection from familiarity in picture recognition. *Cognitive Brain Research*, 15, 191-205.
- Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 33, 2-17.
- Curran, T., DeBuse, C., Woroch, B., & Hirshman, E. (2006). Combined pharmacological and electrophysiological dissociation of familiarity and recollection. *Journal of Neuroscience*, 26, 1979-2009.
- Curran, T., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning Memory and Cognition*, 21, 531-547.
- Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory*. San Diego: Academic Press.
- Dennis, S., & Humphreys, M. S. (1998). Cuing for context: An alternative to global matching models of recognition memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 109-127). Oxford, England: Oxford University Press.
- Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108, 452-477.
- Dennis, S., Lee, M. D., & Kinnell, A. (submitted). Bayesian analysis of recognition memory: The case of the list-length effect.
- Dewhurst, S. A. (2001). Category repetition and false recognition: Effects of instance frequency and category size. *Journal of Memory and Language*, 44, 153-167.

- Dewhurst, S. A., Holmes, S. J., Brandt, K. R., & Dean, G. M. (2006). Measuring the speed of the conscious components of recognition memory: Remembering is faster than knowing. *Consciousness and Cognition*, *15*, 147-162.
- Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Memory and Cognition*, *33*, 1289-1302.
- Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin and Review*, *13*, 1-21.
- Dodson, C. S., Holland, P. W., & Shimamura, A. P. (1998). On the recollection of specific- and partial-source information. *Journal of Experimental Psychology: Learning Memory and Cognition*, *24*, 1121-1136.
- Dodson, C. S., & Johnson, M. K. (1996). Some problems with the process-dissociation approach to memory. *Journal of Experimental Psychology: General*, *125*, 181-194.
- Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory and Cognition*, *24*, 523-533.
- Dorfman, D. D., & Alf, E., Jr. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence interval - rating-method data. *Journal of Mathematical Psychology*, *6*, 487-496.
- Dunn, J. C. (2004). Remember-Know: A matter of confidence. *Psychological Review*, *111*, 524-542.
- Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology*. (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover.
- Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience*, *1*, 41-50.
- Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, *30*, 123-152.
- Ellenbogen, J. M., Hulbert, J. C., Stickgold, R., Dinges, D. F., & Thompson-Schill, S. L. (2006). Interfering with theories of sleep and memory: Sleep, declarative memory, and associative interference. *Current Biology*, *16*, 1290-1294.
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods Instruments and Computers*, *28*, 1-11.

- Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin and Review*, *12*, 403-408.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.
- Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, *23*, 132-138.
- Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379-390.
- Gallo, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning Memory and Cognition*, *30*, 120-128.
- Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, *16*, 309-313.
- Gardiner, J. M., & Gregg, V. H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin and Review*, *4*, 474-479.
- Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 229-244). Oxford: Oxford University Press.
- Gehring, R. E., Toglia, M. P., & Kimble, G. A. (1976). Recognition memory for words and pictures at short and long retention intervals. *Memory & Cognition*, *4*, 256-260.
- Gillund, G., & Shiffrin, R. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1-67.
- Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition*, *13*, 8-20.
- Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*, 5-16.
- Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning & Memory*, *7*, 311-325.

- Gold, J. J., Smith, C. N., Bayley, P. J., Shrager, Y., Brewer, J. B., Stark, C. E. L., et al. (2006). Item memory, source memory, and the medial temporal lobe: Concordant findings from fMRI and memory-impaired patients. *Proceedings-National Academy of Sciences USA*, *103*, 9351-9356.
- Gonsalves, B. D., Kahn, I., Curran, T., Norman, K. A., & Wagner, A. D. (2005). Memory strength and repetition suppression: Multimodal imaging of medial temporal cortical contributions to recognition. *Neuron*, *47*, 751-761.
- Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1355-1369.
- Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 846-858.
- Harvey, L. O., Jr. (2001). *Parameter estimation of signal detection models: RscorePlus user's manual*. Boulder, CO: Author.
- Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*, 1210-1230.
- Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language*, *55*, 495-514.
- Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin and Review*, *10*, 718-723.
- Henson, R. N. A., Cansino, S., Herron, J. E., Robb, W. G. K., & Rugg, M. D. (2003). A familiarity signal in human anterior medial temporal cortex? *Hippocampus*, *13*, 301-304.
- Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin and Review*, *11*, 125-130.
- Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528-551.
- Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory and Cognition*, 336-350.

- Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, 33, 1-18.
- Hintzman, D. L., & Curran, T. (1995). When encoding fails: Instructions, feedback, and registration without learning. *Memory and Cognition*, 23, 213-226.
- Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning Memory and Cognition*, 18, 667-680.
- Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 302-313.
- Hirshman, E., Fisher, J., Henthorn, T., Arndt, J., & Passannante, A. (2002). Midazolam amnesia and dual-process models of the word-frequency mirror effect. *Journal of Memory and Language*, 47, 499-516.
- Hockley, W. E., & Nieuwomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion shifts. *Memory and Cognition*, 35, 679-688.
- Howard, M.W., & Kahana, M.J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, 46, 85-98.
- Howell, D. C. (2002). *Statistical Methods for Psychology* (5th ed.). Belmont, CA: Thomson Wadsworth.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513-541.
- Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General*, 122, 139-154.
- Jones, T. C. (2005). Study repetition and the rejection of conjunction lures. *Memory*, 13, 499-515.
- Joordens, S., & Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 1534-1555.

- Kahana, M. J., Rizzuto, D. S., & Schneider, A. R. (2005). Theoretical correlations and measured correlations: Relating recognition and recall in four distributed memory models. *Journal of Experimental Psychology: Learning Memory and Cognition*, 31, 933-953.
- Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 701-722.
- Kinnell, A., & Dennis, S. (2007). *The list-length effect in recognition memory: An analysis using remember-know responses*. Paper presented at the Presented at the 8th Meeting of the Australasian Society for Cognitive Science.
- Kirwan, C. B., & Stark, C. E. L. (2004). Medial temporal lobe activation during encoding and retrieval of novel face-name pairs. *Hippocampus*, 14, 919-930.
- Kirwan, C. B., & Stark, C. E. L. (2007). Overcoming interference: An fMRI investigation of pattern separation in the medial temporal lobe. *Learning and Memory*, 14, 625-633.
- Kumaran, D., & Maguire, E. A. (2007a). Match-mismatch processes underlie human hippocampal responses to associative novelty. *Journal of Neuroscience*, 27, 8517-8524.
- Kumaran, D., & Maguire, E. A. (2007b). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus*, 17, 735-748.
- Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, 124, 161-180.
- Lamberts, K. (2005). Mathematical modelling of cognition. In K. Lamberts & R. L. Goldstone (Eds.), *Handbook of Cognition* (pp. 407-421). London: Sage.
- Lampinen, J. M., Odegard, T. N., & Neuschatz, J. S. (2004). Robust recollection rejection in the memory conjunction paradigm. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30, 332-342.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- Light, L. L., Chung, C., Pendergrass, R., & Van Ocker, J. C. (2006). Effects of repetition and response deadline on item recognition in young and older adults. *Memory and Cognition*, 34, 335-343.



- Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, 74, 100-109.
- Macho, S. (2004). Modeling associative recognition: A comparison of two-high-threshold, two-high-threshold signal detection, and mixture distribution models. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30, 83-97.
- MacLeod, M. D., & Macrae, C. N. (2001). Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychological Science*, 12, 148-152.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
- Macmillan, N. A., Rotello, C. M., & Miller, J. O. (2004). The sampling distributions of gaussian ROC statistics. *Perception and Psychophysics*, 66, 406-421.
- Macrae, C. N., & Macleod, M. D. (1999). On recollections lost: When practice makes imperfect. *Journal of Personality and Social Psychology*, 77, 463-473.
- Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning Memory and Cognition*, 28, 380-387.
- Malmberg, K. J. (in press). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*.
- Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30, 319-331.
- Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning Memory and Cognition*, 31, 322-336.
- Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin and Review*, 13, 99-105.
- Malmberg, K. J., & Xu, J. (2007). On the flexibility and the fallibility of associative memory. *Memory and Cognition*, 35, 545-556.
- Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic

- recognition memory by Midazolam. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30, 540-549.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, 87, 252-271.
- Manns, J. R., Hopkins, R. O., Reed, J. M., Kitchener, E. G., & Squire, L. R. (2003). Recognition memory and the human hippocampus. *Neuron*, 37, 171-180.
- Mayes, A. R., Holdstock, J. S., Isaac, C. L., Hunkin, N. M., & Roberts, N. (2002). Relative sparing of item recognition memory in a patient with adult-onset damage limited to the hippocampus. *Hippocampus*, 12, 325-340.
- Mayes, A. R., Isaac, C. L., Holdstock, J. S., Hunkin, N. M., Montaldi, D., Downes, J. J., et al. (2001). Memory for single items, word pairs, and temporal order of different kinds in a patient with selective hippocampal lesions. *Cognitive Neuropsychology*, 18, 97-124.
- McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, 105, 724-760.
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419-457.
- Mensink, G. J. M., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, 95, 434-455.
- Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin and Review*, 14, 858-865.
- Montaldi, D., Spencer, T. J., Roberts, N., & Mayes, A. R. (2006). The neural system that mediates familiarity memory. *Hippocampus*, 16, 504-520.
- Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 28, 1095-1110.
- Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64, 482-488.

- Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89, 609-626.
- Murdock, B. B., & Kahana, M. J. (1993a). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19, 689-697.
- Murdock, B. B., & Kahana, M. J. (1993b). List-strength and list-length effects: Reply to Shiffrin, Ratcliff, Murnane, and Nobel (1993). *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19, 1450-1453.
- Murnane, K., & Shiffrin, R. M. (1991a). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 17, 855-874.
- Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory and Cognition*, 19, 119-130.
- Neath, I. (1993). Distinctiveness and serial position effects in recognition. *Memory & Cognition*, 21, 689-698.
- Nelson, T. O. (2003). Relevance of unjustified strong assumptions when utilizing signal detection theory. *Behavioural and Brain Sciences*, 26, 351.
- Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 384-413.
- Norman, K. A. (1999). *Differential effects of list strength on recollection and familiarity*. Unpublished doctoral dissertation, Harvard University.
- Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28, 1083-1094.
- Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110, 611-646.
- Norman, K. A., Tepe, K., Nyhus, E., & Curran, T. (2008). Event-related potential correlates of interference effects on recognition memory. *Psychonomic Bulletin and Review*, 15, 36-43.
- Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning Memory and Cognition*, 14, 700-708.

- O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
- Odegard, T. N., Lampinen, J. M., & Toglia, M. P. (2005). Meaning's moderating effect on recollection rejection. *Journal of Memory and Language*, 53, 416-429.
- Ohrt, D. D., & Gronlund, S. D. (1999). List-length effect and continuous memory: Confounds and solutions. In C. Izawa (Ed.), *On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson-Shiffrin model*. (pp. 105-125). Mahwah, NJ: Erlbaum.
- Park, H., Reder, L. M., & Dickison, D. (2005). The effects of word frequency and similarity on recognition judgments: The role of recollection. *Journal of Experimental Psychology: Learning Memory and Cognition*, 31, 568-578.
- Perfect, T. J., Stark, L. J., Tree, J. J., Moulin, C. J., Ahmed, L., & Hutter, R. (2004). Transfer appropriate forgetting: The cue-dependent nature of retrieval-induced forgetting. *Journal of Memory and Language*, 51, 399-417.
- Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, 83, 304-308.
- Prull, M. W., Dawes, L. L. C., Martin, A. M., Rosenberg, H. F., & Light, L. L. (2006). Recollection and familiarity in recognition memory: Adult age differences and neuropsychological test correlates. *Psychology and Aging*, 21, 107-118.
- Qin, J., Raye, C. L., Johnson, M. K., & Mitchell, K. J. (2001). Source ROCs are (typically) curvilinear: Comment on Yonelinas (1999). *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 1110-1116.
- Raaijmakers, J. G. W., & Shiffrin, R. (1981). Search of associative memory. *Psychological Review*, 88, 93-134.
- Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect 1: Data and discussion. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16, 163-178.
- Ratcliff, R., Gronlund, S. D., & Sheu, C. F. (1992). Testing global memory models using ROC curves. *Psychological Review*, 99, 518-535.
- Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic: Functions and

- implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20, 763-785.
- Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 294-320.
- Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, 181, 574-576.
- Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning Memory and Cognition*, 33, 305-320.
- Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Human Learning & Memory*, 6, 91-105.
- Rotello, C. M., & Heit, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject? *Journal of Memory and Language*, 40, 432-453.
- Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition*, 28, 907-922.
- Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review*, 111, 588-616.
- Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language*, 43, 67-88.
- Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences*, 11, 251-257.
- Rugg, M. D., & Yonelinas, A. P. (2003). Human recognition memory: A cognitive neuroscience perspective. *Trends in Cognitive Sciences*, 7, 313-319.
- Schooler, L. J., & Shiffrin, R. M. (2005). Efficiently measuring recognition performance with sparse data. *Behavior Research Methods Instruments and Computers*, 37, 3-10.

- Schulman, A. I. (1974). The declining course of recognition memory. *Memory and Cognition*, 2, 14-18.
- Shaw, J. S., Bjork, R. A., & Handal, A. (1995). Retrieval-induced forgetting in an eyewitness-memory paradigm. *Psychonomic Bulletin and Review*, 2, 249.
- Sheffert, S. M., & Shiffrin, R. M. (2003). Auditory registration without learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, 29, 10-21.
- Shepard, R. N. (1967). Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6, 156-163.
- Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 267-287.
- Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect 2: Theoretical mechanisms. *Journal of Experimental Psychology: Learning Memory and Cognition*, 16, 179-195.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - Retrieving Effectively from Memory. *Psychonomic Bulletin and Review*, 4, 145-166.
- Simmons, S. G., & Estes, Z. (2006). Using latent semantic analysis to estimate similarity. *Proceedings of the Cognitive Science Society*, 2169-2173.
- Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory and Cognition*, 34, 125-137.
- Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory and Cognition*, 33, 151-170.
- Slotnick, S. D., Klein, S. A., Dodson, C. S., & Shimamura, A. P. (2000). An analysis of signal detection and threshold models of source memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 26, 1499-1517.
- Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: a new perspective. *Nature Reviews Neuroscience*, 8, 872-883.

- Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 29, 760-766.
- Stretch, V., & Wixted, J. T. (1998a). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning Memory and Cognition*, 24, 1397-1410.
- Stretch, V., & Wixted, J. T. (1998b). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 24, 1379-1396.
- Strong, E. K., Jr. (1912). The effect of length of series upon recognition memory. *Psychological Review*, 19, 447-462.
- Strong, E. K., Jr. (1913). The effect of time-interval upon recognition memory. *Psychological Review*, 20, 339-372.
- Toth, J. P. (1996). Conceptual automaticity in recognition memory: Levels-of-processing effects on familiarity. *Canadian Journal of Experimental Psychology*, 50, 123-138.
- Tsivilis, D., Vann, D. S., Denby, C., Roberts, N., Mayes, A. R., Montaldi, D., et al. (2008). A disproportionate role for the fornix and mammillary bodies in recall versus recognition memory. *Nature Neuroscience*, 11, 834-842.
- Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26, 1-12.
- Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free-recall. *Journal of Experimental Psychology* 92(3), 297-304.
- Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, 64, 49-60.
- Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 582-600.
- Verde, M. F. (2004). The retrieval practice effect in associative recognition. *Memory and Cognition*, 32, 1265-1272.
- Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of  $d'$ ,  $Az$ , and  $A'$ . *Perception and Psychophysics*, 68, 643-654.

- Verde, M. F., & Rotello, C. M. (2003). Does familiarity change in the revelation effect? *Journal of Experimental Psychology: Learning Memory and Cognition*, 29, 739-746.
- Verde, M. F., & Rotello, C. M. (2004). Strong memories obscure weak memories in associative recognition. *Psychonomic Bulletin and Review*, 11, 1062-1066.
- Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory and Cognition*, 35, 254-262.
- Weber, E. (1988). Expectation and variance of item resemblance distributions in a convolution-correlation model of distributed memory. *Journal of Mathematical Psychology*, 32, 1-43.
- Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, 54, 277-296.
- Westerman, D. L. (2000). Recollection-based recognition eliminates the revelation effect in memory. *Memory and Cognition*, 28, 167-175.
- Westerman, D. L. (2001). The role of familiarity in item recognition, associative recognition, and plurality recognition on self-paced and speeded tests. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 723-732.
- Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.
- Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, 114, 152-176.
- Wixted, J. T., Ghadisha, H., & Vera, R. (1997). Recall latency following pure- and mixed-strength lists: A direct test of the relative strength model of free recall. *Journal of Experimental Psychology: Learning Memory and Cognition*, 23, 523-538.
- Wixted, J. T., & Squire, L. R. (2004). Recall and recognition are equally impaired in patients with selective hippocampal damage. *Cognitive, Affective, & Behavioral Neuroscience*, 4, 58-66.
- Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of Remember/Know judgments. *Psychonomic Bulletin and Review*, 11, 616-641.
- Xu, J., & Malmberg, K. J. (2007). Modeling the effects of verbal and nonverbal pair strength on associative recognition. *Memory and Cognition*, 35, 526-544.



- Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning Memory and Cognition*, 20, 1341-1354.
- Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory and Cognition*, 25, 747-763.
- Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 1415-1434.
- Yonelinas, A. P. (2001). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 356, 1363-1374.
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46, 441-517.
- Yonelinas, A. P., & Jacoby, L. L. (1994). Dissociations of processes in recognition memory: Effects of interference and of response speed. *Canadian Journal of Experimental Psychology*, 48, 516-535.
- Yonelinas, A. P., Kroll, N. E. A., Dobbins, I., Lazzara, M., & Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember-know, process dissociation, and receiver operating characteristic data. *Neuropsychology*, 12, 323-339.
- Yonelinas, A. P., Murdock, B. B., & Hockley, W. E. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, 345-355.
- Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, 133, 800-832.
- Zaki, S. R., & Nosofsky, R. M. (2001). Exemplar accounts of blending and distinctiveness effects in perceptual old-new recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 1022-1041.

# Appendix 1

In this Appendix, we list raw data (hits and false alarms), equal-variance SDT measures ( $d'$ ,  $c$ ) and unequal variance measures ( $A_z$ ,  $c_a$ ) for each experimental condition of each experiment described in the main body of the thesis.

## Experiment 1

**Table A.1. Hits and false alarms across encoding tasks.**

List type	HR Targets		FAR Unrelated lures	
	<i>M</i>	<i>SEM</i>	<i>M</i>	<i>SEM</i>
<b>(N = 36) Size judgement</b>				
<b>Short</b>	.90	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$	.11	$\begin{smallmatrix} \top & \top \\ n & ** \end{smallmatrix}$
<b>Long</b>	.91	$\begin{smallmatrix} \top & \\ n & \perp \end{smallmatrix}$	.11	$\begin{smallmatrix} \top & \\ ** & \perp \end{smallmatrix}$
<b>Strong</b>	.89	$\begin{smallmatrix} \perp & \perp \end{smallmatrix}$	.07	$\begin{smallmatrix} \perp & \perp \end{smallmatrix}$
<b>(N = 36) Pleasantness judgement</b>				
<b>Short</b>	.95	$\begin{smallmatrix} \top & \top \\ n & * \end{smallmatrix}$	.10	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$
<b>Long</b>	.94	$\begin{smallmatrix} \top & \\ n & \perp \end{smallmatrix}$	.12	$\begin{smallmatrix} \top & \\ * & \perp \end{smallmatrix}$
<b>Strong</b>	.93	$\begin{smallmatrix} \perp & \perp \end{smallmatrix}$	.08	$\begin{smallmatrix} \perp & \perp \end{smallmatrix}$

Note. HR = hits; FAR = false alarms. *n* non-significant; \*  $p < .05$ ; \*\*  $p < .01$ .

**Table A.2. Sensitivity ( $d'$ ) and bias ( $c$ ) across encoding tasks.**

List type	$d'$		$c$	
	<i>M</i>	<i>SEM</i>	<i>M</i>	<i>SEM</i>
<b>(N = 36) Size judgement</b>				
<b>Short</b>	2.76	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$	-.03	$\begin{smallmatrix} \top & \top \\ n & ** \end{smallmatrix}$
<b>Long</b>	2.75	$\begin{smallmatrix} \top & \\ n & \perp \end{smallmatrix}$	-.05	$\begin{smallmatrix} \top & \\ ** & \perp \end{smallmatrix}$
<b>Strong</b>	2.89	$\begin{smallmatrix} \perp & \perp \end{smallmatrix}$	.14	$\begin{smallmatrix} \perp & \perp \end{smallmatrix}$
<b>(N = 36) Pleasantness judgement</b>				
<b>Short</b>	3.19	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$	-.21	$\begin{smallmatrix} \top & \top \\ n & ** \end{smallmatrix}$
<b>Long</b>	3.02	$\begin{smallmatrix} \top & \\ n & \perp \end{smallmatrix}$	-.14	$\begin{smallmatrix} \top & \\ \dagger & \perp \end{smallmatrix}$
<b>Strong</b>	3.12	$\begin{smallmatrix} \perp & \perp \end{smallmatrix}$	.05	$\begin{smallmatrix} \perp & \perp \end{smallmatrix}$

Note.  $d'$  = sensitivity,  $c$  = bias. *n* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ .

**Table A.3. Sensitivity ( $A_z$ ) and bias ( $c_a$ ) across encoding tasks.**

List type	$A_z$			$C_a$			
	$M$		$SEM$	$M$		$SEM$	
(N = 34)	Size judgement						
Short	.94	$\tau$ $n$	$\tau$ $n$	.01	0.12	$\tau$ $n$ *	0.05
Long	.94	$\tau$ $n$	$\perp$ $n$	.01	0.09	$\tau$ $\perp$	0.04
Strong	.95	$\perp$ $n$	$\perp$ $n$	.01	0.23	** $\perp$	0.05
(N = 34)	Pleasantness judgment						
Short	.97	$\tau$ $n$	$\tau$ $n$	.01	-.01	$\tau$ $n$ *	0.05
Long	.96	$\tau$ $n$	$\perp$ $n$	.01	0.05	$\tau$ $\perp$	0.04
Strong	.96	$\perp$ $n$	$\perp$ $n$	.01	0.10	$\perp$ $n$	0.05

Note.  $A_z$  = area under the ROC;  $c_a$  = response bias (hits and false alarms from  $X_3$ , which separates *guess old* from *guess new* responses).  $n$  non-significant; \*  $p < .05$ ; \*\*  $p < .01$ .

## Experiment 2

Table A.4. Hits and false alarms.

List type	HR Targets			FAR SP lures			FAR Unrelated		
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>
Short	.80	⌈ n ⊥	⌈ * ⊥	.01	.45	⌈ n ⊥	.02	.10	⌈ * ⊥
Long	.81	⌈ *	⊥	.01	.46	⌈ n ⊥	.02	.12	⌈ * ⊥
Strong	.78	⊥	⊥	.01	.45	⊥	.02	.10	⊥

Note. SP lures = switched-plurality lures; *n* non-significant; \*  $p \leq .05$ ; \*\*  $p < .01$ .  $N = 126$ .

Table A.5. Sensitivity ( $d'$ ) and bias ( $c$ ).

List type	$d'$			$c$		
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>
Short	1.63	⌈ n ⊥	0.05	-.13	⌈ n ⊥	0.03
Long	1.61	⌈ *	0.05	-.16	⌈ *	0.03
Strong	1.57	⊥	0.05	-.07	⊥	0.03

Note.  $d'$  = sensitivity,  $c$  = bias. *n* non-significant; \*  $p < .05$ ; \*\*  $p < .01$ ;  $N = 126$ .

Table A.6. Sensitivity ( $d'$ ) for related and unrelated lures.

List type	$d'$ (SP lures)			$d'$ (unrelated lures)		
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>
Short	1.08	⌈ n ⊥	0.07	2.36	⌈ n ⊥	0.07
Long	1.09	⌈ *	0.06	2.32	⌈ *	0.06
Strong	1.01	⊥	0.06	2.34	⊥	0.07

Note. SP lures = switched-plurality lures; *n* non-significant;  $N = 126$ .

### Experiment 3

Table A.7. Hits and false alarms across encoding tasks and lure types.

List type	HR Targets		FAR SP lures		FAR Unrelated				
	<i>M</i>	<i>SEM</i>	<i>M</i>	<i>SEM</i>	<i>M</i>	<i>SEM</i>			
(N = 66) Size judgement task									
Short	.79	$\begin{smallmatrix} \top & \top \\ \dagger & \dagger \end{smallmatrix}$	.02	.43	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$	.03	.10	$\begin{smallmatrix} \top & \top \\ n & * \end{smallmatrix}$	.01
Long	.82	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$ ***	.01	.44	$\begin{smallmatrix} \top & \top \\ \perp & n \end{smallmatrix}$	.03	.11	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$ **	.01
Strong	.75	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$	.02	.44	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$	.02	.08	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$ **	.01
(N = 66) Pleasantness judgement task									
Short	.82	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$	.02	.50	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$	.03	.11	$\begin{smallmatrix} \top & \top \\ * & \top \end{smallmatrix}$ **	.01
Long	.80	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$	.02	.49	$\begin{smallmatrix} \top & \top \\ \perp & n \end{smallmatrix}$	.03	.10	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$ **	.01
Strong	.80	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$	.02	.52	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$	.03	.07	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$ **	.01

Note. SP = switched plurality; *n* non-significant;  $\dagger p < .10$ ;  $* p \leq .05$ ;  $** p < .01$ ;  $*** p < .001$ .

Table A.8. Sensitivity ( $d'$ ) and bias ( $c$ ) across encoding tasks.

List type	<i>d'</i>		<i>c</i>		
	<i>M</i>	<i>SEM</i>	<i>M</i>	<i>SEM</i>	
(N = 66) Size judgment					
Short	1.63	$\begin{smallmatrix} \top & \top \\ * & n \end{smallmatrix}$	0.08	$\begin{smallmatrix} -0.07 & \\ \top & \top \\ \dagger & * \end{smallmatrix}$	0.04
Long	1.70	$\begin{smallmatrix} \top & \top \\ n & \perp \end{smallmatrix}$	0.08	$\begin{smallmatrix} -0.15 & \\ \top & \top \\ \perp & \perp \end{smallmatrix}$	0.05
Strong	1.51	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$	0.08	$\begin{smallmatrix} 0.02 & \\ \perp & \perp \\ \perp & \perp \end{smallmatrix}$	0.04
(N = 66) Pleasantness judgment					
Short	1.67	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$	0.08	$\begin{smallmatrix} -0.20 & \\ \top & \top \\ \dagger & \dagger \end{smallmatrix}$	0.04
Long	1.62	$\begin{smallmatrix} \top & \top \\ \perp & n \end{smallmatrix}$	0.08	$\begin{smallmatrix} -0.13 & \\ \top & \top \\ \perp & \perp \end{smallmatrix}$	0.05
Strong	1.70	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$	0.08	$\begin{smallmatrix} -0.13 & \\ \perp & \perp \\ \perp & \perp \end{smallmatrix}$	0.04
(N = 132) Size and pleasantness					
Short	1.66	$\begin{smallmatrix} \top & \top \\ n & n \end{smallmatrix}$	0.06	$\begin{smallmatrix} -0.13 & \\ \top & \top \\ n & ** \end{smallmatrix}$	0.03
Long	1.66	$\begin{smallmatrix} \top & \top \\ \perp & n \end{smallmatrix}$	0.06	$\begin{smallmatrix} -0.14 & \\ \top & \top \\ \perp & \perp \end{smallmatrix}$	0.03
Strong	1.60	$\begin{smallmatrix} \top & \top \\ \perp & \perp \end{smallmatrix}$	0.06	$\begin{smallmatrix} -0.05 & \\ \perp & \perp \\ \perp & \perp \end{smallmatrix}$	0.03

Note.  $d'$  = sensitivity,  $c$  = bias. *n* non-significant;  $\dagger p < .10$ ;  $* p < .05$ ;  $\dagger p < .01$ .

**Table A.9. Sensitivity ( $d'$ ) for related and unrelated lures.**

List type	$d'$ (SP lures)		$d'$ (unr. lures)	
	$M$	$SEM$	$M$	$SEM$
<b>(<math>N = 66</math>) Size judgment</b>				
Short	1.10	$\tau$ $\tau$ $n$	0.10	2.32 $\tau$ $\tau$ $n$
Long	1.15	$\tau$ $\perp$ $*$	0.10	2.43 $\tau$ $\perp$ $n$
Strong	0.91	$\perp$ $\perp$	0.10	2.32 $\perp$ $\perp$
<b>(<math>N = 66</math>) Pleasantness judgment</b>				
Short	1.03	$\tau$ $\tau$ $n$	0.10	2.40 $\tau$ $\tau$ $n$
Long	0.97	$\tau$ $\perp$ $n$	0.10	2.33 $\tau$ $\perp$ $n$
Strong	0.92	$\perp$ $\perp$	0.10	2.56 $\perp$ $\perp$
<b>(<math>N = 132</math>) Size and Pleasantness</b>				
Short	1.06	$\tau$ $\tau$ $n$	0.07	2.36 $\tau$ $\tau$ $n$
Long	1.06	$\tau$ $\perp$ $*$	0.07	2.38 $\tau$ $\perp$ $n$
Strong	0.92	$\perp$ $\perp$	0.07	2.44 $\perp$ $\perp$

Note. unr. = unrelated.  $n$  non-significant;  $\dagger p < .10$ ;  $* p \leq .05$ .

**Table A.10. Sensitivity ( $A_z$ ) across encoding tasks and comparison types.**

List type	Size		Pleasantness	
	$M$	$SEM$	$M$	$SEM$
<b>Studied vs. unrelated lures</b>				
Short	.90	$\tau$ $\tau$ $n$	.01	.91 $\tau$ $\tau$ $n$
Long	.92	$\tau$ $\perp$ $n$	.01	.91 $\tau$ $\perp$ $n$
Strong	.91	$\perp$ $\perp$	.01	.92 $\perp$ $\perp$
<b>Studied vs. switched plurality</b>				
Short	.75	$\tau$ $\tau$ $n$	.02	.74 $\tau$ $\tau$ $n$
Long	.75	$\tau$ $\perp$ $*$	.02	.74 $\tau$ $\perp$ $*$
Strong	.70	$\perp$ $\perp$	.01	.70 $\perp$ $\perp$
<b>Switched plurality vs. unrelated</b>				
Short	.69	$\tau$ $\tau$ $n$	.02	.74 $\tau$ $\tau$ $n$
Long	.72	$\tau$ $\perp$ $**$	.02	.70 $\tau$ $\perp$ $**$
Strong	.80	$\perp$ $\perp$	.01	.79 $\perp$ $\perp$

Note.  $A_z$  = estimate of the area under the ROC;  $n$  non-significant;  $* p < .05$ ;  $** p < .01$ . Size = size judgement task; Pleasantness = pleasantness judgement task.  $N = 119$ .

**Table A.11. Bias ( $c_a$ ) across encoding tasks and comparison types.**

List type	Size				Pleasantness			
	$M$			$SEM$	$M$			$SEM$
<b>Studied vs. unrelated lures</b>								
<b>Short</b>	0.23		$\top$	$\top$	0.04		$\top$	$\top$
			$n$				$n$	
<b>Long</b>	0.20	$\top$	$\perp$	**	0.04	0.21	$\top$	$\perp$
		**					$n$	*
<b>Strong</b>	0.38	$\perp$		$\perp$	0.04	0.26	$\perp$	$\perp$
<b>Studied vs. switched plurality</b>								
<b>Short</b>	-.29		$\top$	$\top$	0.05	-.48	$\top$	$\top$
			$n$				$n$	
<b>Long</b>	-.36	$\top$	$\perp$	$n$	0.06	-.41	$\top$	$\perp$
		**					$n$	$n$
<b>Strong</b>	-.24	$\perp$		$\perp$	0.05	-.49	$\perp$	$\perp$
<b>Switched plurality vs. unrelated</b>								
<b>Short</b>	0.71		$\top$	$\top$	0.06	0.57	$\top$	
			$\dagger$				$n$	$\top$
<b>Long</b>	0.66	$\top$	$\perp$	$n$	0.06	0.61	$\top$	$\perp$
							$\dagger$	
<b>Strong</b>	0.75	$n$	$\perp$	$\perp$	0.06	0.66	$n$	$\perp$
		$\perp$					$\perp$	

Note.  $c_a$  = response bias (from  $X_3$ );  $n$  non-significant;  $\dagger p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ .

## Experiment 4

Table A.12. Hits and false alarms across retention intervals and lure types.

List type	HR Targets				FAR Switched plurality				FAR Unrelated lures			
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>	
<b>(N = 54) Short retention interval</b>												
Short	.78	⌈	⌈	.02	.45	⌈	⌈	.03	.12	⌈	⌈	.01
		n	**			n	n			n		
Long	.76	⌈	⌊	.02	.49	⌈	⌊	.03	.13	⌈	⌊	.01
		†				*				***		
Strong	.72	⌊	⌊	.02	.41	⌊	⌊	.03	.07	⌊	⌊	.01
<b>(N = 54) Long retention interval</b>												
Short	.78	⌈	⌈	.02	.43	⌈	⌈	.03	.13	⌈	⌈	.01
		n	*			n	n			n		
Long	.77	⌈	⌊	.02	.48	⌈	⌊	.03	.15	⌈	⌊	.02
		†				*				***		
Strong	.73	⌊	⌊	.02	.42	⌊	⌊	.02	.09	⌊	⌊	.01

Note.  $A_z$  = area under the ROC; *n* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p \leq .001$ .

Table A.13. Sensitivity ( $d'$ ) and bias ( $c$ ) across retention intervals.

List type	$d'$				$c$			
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>	
<b>(N = 54) Short retention interval</b>								
Short	1.66	⌈	⌈	0.09	-.05	⌈	⌈	0.05
		†	n			n		
Long	1.51	⌈	⌊	0.09	-.02	⌈	⌊	0.05
		n				***		
Strong	1.62	⌊	⌊	0.09	0.18	⌊	⌊	0.04
<b>(N = 54) Long retention interval</b>								
Short	1.68	⌈	⌈	0.09	-.02	⌈	⌈	0.05
		n	n			n	*	
Long	1.55	⌈	⌊	0.09	-.05	⌈	⌊	0.05
		n				**		
Strong	1.64	⌊	⌊	0.09	0.11	⌊	⌊	0.04
<b>(N = 108) Short and Long interval</b>								
Short	1.67	⌈	⌈	0.06	-.03	⌈	⌈	0.03
		*	n			n		
Long	1.53	⌈	⌊	0.07	-.04	⌈	⌊	0.03
		n				***		
Strong	1.63	⌊	⌊	0.07	0.14	⌊	⌊	0.03

Note.  $d'$  = sensitivity,  $c$  = bias. *n* non-significant; †  $p \leq .1$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .



**Table A.14. Sensitivity ( $d'$ ) for related and unrelated lures.**

List type	$d'$ (SP lures)			$d'$ (unr. lures)		
	$M$		$SEM$	$M$		$SEM$
<b>(<math>N = 54</math>) Short retention interval</b>						
Short	1.01	$\tau$	$\tau$ 0.10	2.16	$\tau$	$\tau$ 0.10
Long	0.82	$\tau$	$\perp$ <sup>n</sup> 0.12	2.04	$\tau$	$\perp$ <sup>n</sup> ** 0.09
Strong	0.89	$\perp$	$\perp$ 0.11	2.26	$\perp$	$\perp$ 0.10
<b>(<math>N = 54</math>) Long retention interval</b>						
Short	1.08	$\tau$	$\tau$ 0.10	2.13	$\tau$	$\tau$ 0.10
Long	0.87	$\tau$	$\perp$ <sup>n</sup> 0.12	2.02	$\tau$	$\perp$ <sup>n</sup> 0.09
Strong	0.95	$\perp$	$\perp$ 0.11	2.21	$\perp$	$\perp$ 0.11
<b>(<math>N = 108</math>) Short and long interval</b>						
Short	1.05	$\tau$	$\tau$ 0.07	2.15	$\tau$	$\tau$ 0.07
Long	0.85	$\tau$	$\perp$ <sup>*</sup> $\dagger$ 0.08	2.03	$\tau$	$\perp$ <sup>n</sup> 0.07
Strong	0.92	$\perp$	$\perp$ 0.08	2.23	$\perp$	$\perp$ 0.07

Note. unr. = unrelated;  $n$  non-significant;  $\dagger p < .10$ ;  $* p < .05$ ;  $** p < .01$ .

**Table A.15. Sensitivity ( $A_z$ ) across retention intervals and comparison types.**

List type	Short interval ( $N = 49$ )			Long interval ( $N = 51$ )		
	$M$		$SEM$	$M$		$SEM$
<b>Studied vs. unrelated lures</b>						
Short	.88	$\tau$	$\tau$ .01	.87	$\tau$	$\tau$ .01
Long	.88	$\tau$	$\perp$ <sup>n</sup> .01	.86	$\tau$	$\perp$ <sup>n</sup> .01
Strong	.89	$\perp$	$\perp$ .01	.88	$\perp$	$\perp$ .01
<b>Studied vs. switched plurality</b>						
Short	.72	$\tau$	$\tau$ .02	.75	$\tau$	$\tau$ .02
Long	.68	$\tau$	$\perp$ <sup>*</sup> $\dagger$ .02	.70	$\tau$	$\perp$ <sup>**</sup> $\dagger$ .02
Strong	.70	$\perp$	$\perp$ .02	.71	$\perp$	$\perp$ .02
<b>Switched plurality vs. unrelated</b>						
Short	.70	$\tau$	$\tau$ .03	.65	$\tau$	$\tau$ .03
Long	.75	$\tau$	$\perp$ <sup>*</sup> .02	.68	$\tau$	$\perp$ <sup>n</sup> ** .02
Strong	.77	$\perp$	$\perp$ .02	.73	$\perp$	$\perp$ .02

Note.  $A_z$  = area under the ROC;  $n$  non-significant;  $\dagger p < .10$ ;  $* p < .05$ ;  $** p < .01$ .

**Table A.16. Bias ( $c_a$ ) across retention intervals and comparison types.**

List type	Short interval ( <i>N</i> = 49)				Long interval ( <i>N</i> = 51)					
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>			
Studied vs. unrelated lures										
Short	0.16		†	†	0.04	0.16		†	†	0.04
Long	0.24	†	†	***	0.05	0.17	†	†	†	0.04
Strong	0.46	†		†	0.04	0.36	†		†	0.04
Studied vs. switched plurality										
Short	-.34		†	†	0.05	-.27		†	†	0.05
Long	-.30	†	†	**	0.05	-.33	†	†	†	0.05
Strong	-.15	†		†	0.05	-.19	†		†	0.05
Switched plurality vs. unrelated										
Short	0.56		†	†	0.05	0.62		†	†	0.05
Long	0.58	†	†	***	0.06	0.53	†	†	†	0.06
Strong	0.82	†		†	0.05	0.72	†		†	0.05

Note.  $c_a$  = response bias;  $n$  non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .

## Experiment 5a

Table A.17. Hits and false alarms across retention intervals.

A items							
List type	HR Targets			FAR SP lures			
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	
Short retention interval							
Short	.74	⌈	⌈	.02	.33	⌈	⌈
		†	**			n	n
Long	.70	⌈	⌊	.02	.29	⌈	⌊
		n				n	
Strong	.68	⌊	⌊	.02	.30	⌊	⌊
Long retention interval							
Short	.70	⌈	⌈	.02	.34	⌈	⌈
		†	†			n	n
Long	.69	⌈	⌊	.02	.35	⌈	⌊
		n				n	
Strong	.66	⌊	⌊	.02	.33	⌊	⌊
B items							
List type	HR Targets			FAR SP lures			
	<i>M</i>		<i>SEM</i>	<i>M</i>		<i>SEM</i>	
Short retention interval							
Short	.68	⌈	⌈	.03	.32	⌈	⌈
		n	***			n	n
Long	.71	⌈	⌊	.03	.32	⌈	⌊
		***				†	
Strong	.84	⌊	⌊	.02	.28	⌊	⌊
Long retention interval							
Short	.71	⌈	⌈	.02	.34	⌈	⌈
		n	***			n	n
Long	.71	⌈	⌊	.02	.34	⌈	⌊
		***				n	
Strong	.82	⌊	⌊	.02	.33	⌊	⌊

Note. HR = hits; FAR = false alarms. SP = switched-plurality. A/B items = items from the beginning of the study list (in *strong* lists, B items are repeated). *n* non-significant; †  $p < .10$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ . Short retention interval = 10 s; Long interval = 120 s;  $N = 48$ .

**Table A.18. Sensitivity ( $d'$ ) and bias ( $c$ ) across item types.**

A items							
List type	$d'$			$c$			
	$M$	$SEM$		$M$	$SEM$		
Short retention interval							
Short	1.24	$\tau$ $n$	$\tau$ $n$	0.12	-.10	$\tau$ $n$	0.06
Long	1.21	$\tau$ $n$	$\perp$ $n$	0.10	0.04	$\tau$ $n$	0.05
Strong	1.11	$\perp$ $n$	$\perp$	0.09	0.03	$\perp$	0.06
Long retention interval							
Short	1.07	$\tau$ $n$	$\tau$ $n$	0.12	-.06	$\tau$ $n$	0.06
Long	0.98	$\tau$ $n$	$\perp$ $n$	0.10	-.07	$\tau$ $n$	0.05
Strong	0.95	$\perp$ $n$	$\perp$	0.10	0.00	$\perp$	0.05
B items							
List type	$d'$			$c$			
	$M$	$SEM$		$M$	$SEM$		
Short retention interval							
Short	1.10	$\tau$ $n$	$\tau$ $n$	0.13	-.05	$\tau$ $n$	0.07
Long	1.19	$\tau$ $***$	$\perp$ $***$	0.11	-.07	$\tau$ $*$	0.07
Strong	1.85	$\perp$ $***$	$\perp$	0.14	-.24	$\perp$	0.05
Long retention interval							
Short	1.11	$\tau$ $n$	$\tau$ $n$	0.10	0.00	$\tau$ $n$	0.06
Long	1.10	$\tau$ $***$	$\perp$ $***$	0.10	-.03	$\tau$ $*$	0.06
Strong	1.56	$\perp$ $***$	$\perp$	0.14	-.21	$\perp$	0.06

Note.  $d'$  = sensitivity,  $c$  = bias.  $n$  non-significant;  $\dagger p < .10$ ;  $* p < .05$ ;  $** p < .01$ ;  $*** p \leq .001$ .  $N = 48$ .

**Table A.19. Sensitivity ( $A_z$ ) across retention intervals and item types.**

List type	Short interval			Long interval		
	$M$		$SEM$	$M$		$SEM$
<b>A items</b>						
Short	.75	$\tau$	$\tau$	.73	$\tau$	$\tau$
		$n$	$n$		$n$	$n$
Long	.76	$\tau$	$\perp$	.73	$\tau$	$\perp$
		$n$			$n$	
Strong	.75	$\perp$	$\perp$	.71	$\perp$	$\perp$
<b>B items</b>						
Short	.74	$\tau$	$\tau$	.74	$\tau$	$\tau$
		$n$			$n$	
Long	.75	$\tau$	$\perp$ ***	.74	$\tau$	$\perp$ ***
		***			***	
Strong	.85	$\perp$	$\perp$	.81	$\perp$	$\perp$

Note.  $A_z$  = area under the ROC; A/B items = early items in study list (in *strong* lists, B items are repeated).  $n$  non-significant; \*\*\*  $p \leq .001$ .  $N = 48$ .

**Table A.20. Bias ( $c_a$ ) across retention intervals and item types.**

List type	Short interval			Long interval		
	$M$		$SEM$	$M$		$SEM$
<b>A items</b>						
Short	-.07	$\tau$	$\tau$	-.05	$\tau$	$\tau$
		$\dagger$	$n$		$n$	$n$
Long	0.04	$\tau$	$\perp$	-.06	$\tau$	$\perp$
		$n$			$\dagger$	
Strong	0.02	$\perp$	$\perp$	0.01	$\perp$	$\perp$
<b>B items</b>						
Short	0.03	$\tau$	$\tau$	-.06	$\tau$	$\tau$
		$n$	**		$n$	**
Long	-.03	$\tau$	$\perp$	-.04	$\tau$	$\perp$
		*			**	
Strong	-.17	$\perp$	$\perp$	-.21	$\perp$	$\perp$

Note.  $c_a$  = bias (hits and false alarms at  $X_3$ ). A/B items = early items in study list (in *strong* lists, B items are repeated).  $n$  non-significant;  $\dagger p < .10$ ; \*  $p < .05$ ; \*\*  $p \leq .01$ .

## Experiment 5b

Table A.21. Hits and false alarms across retention intervals.

A items							
List type	HR Targets			FAR SP lures			
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>
Short retention interval							
Short	.78	⊥	⊥	.03	.33	⊥	.04
		†	*			n	
Long	.72	⊥	⊥	.03	.33	⊥	.03
		n				n	
Strong	.69	⊥	⊥	.04	.34	⊥	.05
Long retention interval							
Short	.78	⊥	⊥	.02	.31	⊥	.02
		*	n			n	
Long	.71	⊥	⊥	.03	.33	⊥	.03
		n				n	
Strong	.73	⊥	⊥	.03	.34	⊥	.03
B items							
List type	HR Targets			FAR SP lures			
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>
Short retention interval							
Short	.73	⊥	⊥	.04	.32	⊥	.03
		n	**			n	*
Long	.71	⊥	⊥	.03	.36	⊥	.03
		⊥				⊥	
Strong	.85	⊥	⊥	.03	.23	⊥	.03
		***				**	
Long retention interval							
Short	.73	⊥	⊥	.03	.34	⊥	.03
		n	**			n	
Long	.69	⊥	⊥	.02	.36	⊥	.04
		⊥				⊥	
Strong	.86	⊥	⊥	.02	.30	⊥	.04
		***				n	

Note. HR = hits; FAR = false alarms. SP = switched-plurality. A/B items = early items in study list (in *strong* lists, *B* items are repeated). *n* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ . Short = 10 s; long interval = 120 s;  $N = 24$ .

Table A.22. Sensitivity ( $d'$ ) and bias ( $c$ ) across item types.

A items							
List type	$d'$			$c$			
	$M$	$SEM$		$M$	$SEM$		
Short retention interval							
Short	1.33	$\tau$ $n$	$\tau$ *	0.16	-0.19	$\tau$ $n$	0.08
Long	1.14	$\tau$ $n$	$\perp$	0.17	-0.08	$\tau$ $n$	0.06
Strong	1.08	$\perp$	$\perp$	0.21	-0.05	$\perp$	0.08
Long retention interval							
Short	1.37	$\tau$ *	$\tau$ $n$	0.13	-0.15	$\tau$ $n$	0.05
Long	1.05	$\tau$ $n$	$\perp$	0.11	-0.07	$\tau$ $n$	0.06
Strong	1.14	$\perp$	$\perp$	0.12	-0.12	$\perp$	0.07
B items							
List type	$d'$			$c$			
	$M$	$SEM$		$M$	$SEM$		
Short retention interval							
Short	1.22	$\tau$ $n$	$\tau$	0.17	-0.11	$\tau$ $n$	0.07
Long	1.02	$\tau$ ***	$\perp$ ***	0.16	-0.09	$\tau$ $n$	0.06
Strong	2.03	$\perp$	$\perp$	0.17	-0.15	$\perp$	0.08
Long retention interval							
Short	1.15	$\tau$ $n$	$\tau$	0.12	-0.12	$\tau$ $n$	0.09
Long	0.93	$\tau$ ***	$\perp$ ***	0.15	-0.05	$\tau$ *	0.07
Strong	1.83	$\perp$	$\perp$	0.19	-0.27	$\perp$	0.08

Note.  $d'$  = sensitivity,  $c$  = bias.  $n$  non-significant; \*  $p < .05$ ; \*\*\*  $p < .001$ .  $N = 24$ .

**Table A.23. Sensitivity ( $A_z$ ) across retention intervals and item types.**

List type	Short interval			Long interval		
	$M$		$SEM$	$M$		$SEM$
<b>A items</b>						
Short	.78	$\tau$	$\tau$	.82	$\tau$	$\tau$
		$n$	*		*	*
Long	.76	$\tau$	$\perp$	.75	$\tau$	$\perp$
		$n$			$n$	
Strong	.72	$\perp$	$\perp$	.74	$\perp$	$\perp$
<b>B items</b>						
Short	.77	$\tau$	$\tau$	.76	$\tau$	$\tau$
		$n$	**		$n$	$n$
Long	.73	$\tau$	$\perp$	.72	$\tau$	$\perp$
		**			***	
Strong	.83	$\perp$	$\perp$	.83	$\perp$	$\perp$

Note.  $A_z$  = area under the ROC; A/B items = items from beginning of study list (in *strong* lists, *B* items are repeated).  $n$  non-significant; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .  $N = 22$ .

**Table A.24. Bias ( $c_a$ ) across retention intervals and item types.**

List type	Short interval			Long interval		
	$M$		$SEM$	$M$		$SEM$
<b>A items</b>						
Short	-.17	$\tau$	$\tau$	-.16	$\tau$	$\tau$
		$n$	$n$		$n$	$n$
Long	-.07	$\tau$	$\perp$	-.08	$\tau$	$\perp$
		$n$			$n$	
Strong	-.05	$\perp$	$\perp$	-.12	$\perp$	$\perp$
<b>B items</b>						
Short	-.11	$\tau$	$\tau$	-.07	$\tau$	$\tau$
		$n$	$n$		$n$	$n$
Long	-.07	$\tau$	$\perp$	-.04	$\tau$	$\perp$
		$n$			$\dagger$	
Strong	-.13	$\perp$	$\perp$	-.19	$\perp$	$\perp$

Note.  $c_a$  = bias (hits and false alarms at  $X_3$ ). A/B items = items from beginning of study list (in *strong* lists, *B* items are repeated).  $n$  non-significant;  $\dagger p < .10$ .



## Experiment 6

Table A.25. Hits and false alarms across retention intervals and item types.

A items								
List type	HR Targets				FAR SP lures			
	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>
Short retention interval								
Short	.76		⌈	.02	.34	⌈ <sub>n</sub>	⌈ <sub>n</sub>	.03
Long	.70	⌈	⊥ ***	.02	.36	⌈	⊥ <sub>n</sub>	.03
Strong	.64	*	⊥	.03	.29	**	⊥	.03
Long retention interval								
Short	.77		⌈ <sub>n</sub>	.02	.38	⌈ <sub>n</sub>	⌈ <sub>n</sub>	.04
Long	.76	⌈	⊥ *	.02	.41	⌈	⊥ <sub>n</sub>	.03
Strong	.71	†	⊥	.02	.34	*	⊥	.03
B items								
List type	HR Targets				FAR SP lures			
	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>
Short retention interval								
Short	.74		⌈ <sub>n</sub>	.03	.37	⌈ <sub>n</sub>	⌈ <sub>n</sub>	.03
Long	.73	⌈	⊥ ***	.02	.39	⌈	⊥ **	.03
Strong	.91	***	⊥	.01	.28	***	⊥	.04
Long retention interval								
Short	.72		⌈ <sub>n</sub>	.03	.32	⌈ <sub>n</sub>	⌈ <sub>n</sub>	.03
Long	.73	⌈	⊥ ***	.03	.37	⌈	⊥ <sub>n</sub>	.03
Strong	.90	***	⊥	.01	.30	*	⊥	.04

Note. HR = hits; FAR = false alarms. A/B items = items from beginning of study list (B items are repeated in *strong* lists). *n* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p \leq .01$ ; \*\*\*  $p < .001$ . Short retention interval = 10 s; Long retention interval = 120 s.  $N = 48$ .

Table A.26. Sensitivity ( $d'$ ) and bias ( $c$ ) across item types.

A items								
List type	$d'$				$c$			
	$M$	$SEM$			$M$	$SEM$		
Short retention interval								
Short	1.28	$\tau$ *	$\tau$ †	0.12	-0.12	$\tau$ n	$\tau$ **	0.06
Long	1.03	$\tau$ n	$\perp$	0.11	-0.09	$\tau$ **	$\perp$	0.07
Strong	1.05	$\perp$	$\perp$	0.13	0.11	$\perp$	$\perp$	0.05
Long retention interval								
Short	1.22	$\tau$ n	$\tau$ n	0.14	-0.20	$\tau$ n	$\tau$ †	0.07
Long	1.02	$\tau$ n	$\perp$	0.11	-0.25	$\tau$ **	$\perp$	0.05
Strong	1.16	$\perp$	$\perp$	0.15	-0.07	$\perp$	$\perp$	0.06
B items								
List type	$d'$				$c$			
	$M$	$SEM$			$M$	$SEM$		
Short retention interval								
Short	1.13	$\tau$ n	$\tau$	0.12	-0.15	$\tau$ n	$\tau$ **	0.06
Long	0.99	$\tau$ ***	$\perp$ ***	0.11	-0.18	$\tau$ **	$\perp$	0.05
Strong	2.16	$\perp$	$\perp$	0.16	-0.38	$\perp$	$\perp$	0.07
Long retention interval								
Short	1.24	$\tau$ n	$\tau$	0.14	-0.04	$\tau$ †	$\tau$	0.07
Long	1.11	$\tau$ ***	$\perp$ ***	0.13	-0.16	$\tau$ **	$\perp$ ***	0.07
Strong	1.97	$\perp$	$\perp$	0.17	-0.40	$\perp$	$\perp$	0.07

Note. A = study items; B = interference items (repeated in *strong* lists). Short retention interval = 10 s; Long retention interval = 120 s.  $n$  non-significant;  $\dagger p < .10$ ;  $* p \leq .05$ ;  $** p < .01$ ;  $*** p < .001$ .  $N = 48$ .

**Table A.27. Sensitivity ( $A_z$ ) across retention intervals and item types.**

List type	A items				B items			
	<i>M</i>	<i>(N = 46)</i>		<i>SEM</i>	<i>M</i>	<i>(N = 45)</i>		<i>SEM</i>
Short retention interval								
Short	.78	τ	τ	.02	.74	τ	τ	.03
		†	*			n		
Long	.74	τ	⊥	.02	.71	τ	⊥	***
		n				***		
Strong	.73	⊥	⊥	.02	.89	⊥	⊥	.02
Long retention interval								
Short	.75	τ	τ	.03	.78	τ	τ	.02
		n	n			*		
Long	.72	τ	⊥	.02	.73	τ	⊥	***
		n				***		
Strong	.74	⊥	⊥	.02	.87	⊥	⊥	.02

*Note.*  $A_z$  = area under the ROC; A/B items = items from beginning of study list (*B* items are repeated in *strong* lists). *n* non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*\*  $p < .001$ . Short retention interval = 10 s; Long retention interval = 120 s.

**Table A.28. Bias ( $c_a$ ) across retention intervals and item types.**

List type	A items			B items			
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>
Short retention interval							
Short	-0.11		$\tau$ $\tau$	0.06	-0.12	$\tau$ $\tau$	0.07
			$n$ **			$n$ *	
Long	-0.09	$\tau$	$\perp$	0.07	-0.16	$\tau$ $\perp$	0.05
		**				*	
Strong	0.13	$\perp$	$\perp$	0.05	-0.29	$\perp$	$\perp$
Long retention interval							
Short	-0.17		$\tau$ $\tau$	0.06	0.02	$\tau$ $\tau$	0.06
			$n$ *			*	
Long	-0.20	$\tau$	$\perp$	0.04	-0.11	$\tau$ $\perp$	***
		**				*	
Strong	-0.04	$\perp$	$\perp$	0.05	-0.26	$\perp$	$\perp$

*Note.*  $c_a$  = bias (hits and false alarms for criterion separating *guess old* from *guess new* responses). A = targets; B = interference items (*B* items are repeated in *strong* lists). *n* non-significant; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ . Short retention interval = 10 s; Long retention interval = 120 s.

## Experiment 7

Table A.29. Hits and false alarms across encoding tasks and lure types.

List type	HR Targets				FAR SP lures				FAR Unrelated						
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>				
Short retention interval															
Short	.76		τ	τ	.01	.41		τ	τ	.02	.13		τ	τ	.01
			n					n					n		
Long	.75	τ	⊥	***	.02	.42	τ	⊥	***	.02	.15	τ	⊥	***	.01
		***					***					***			
Strong	.64	⊥		⊥	.02	.34	⊥		⊥	.02	.08	⊥		⊥	.01
Long retention interval															
Short	.75		τ	τ	.02	.39		τ	τ	.02	.15		τ	τ	.01
			n					n					n		
Long	.73	τ	⊥	***	.02	.40	τ	⊥	***	.02	.15	τ	⊥	***	.01
		**					**					**			
Strong	.68	⊥		⊥	.02	.34	⊥		⊥	.02	.11	⊥		⊥	.01

Note. SP = switched plurality; *n* non-significant; \*  $p \leq .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .  $N = 96$ .

Table A.30. Sensitivity ( $d'$ ) and bias ( $c$ ) across retention intervals.

List type	<i>d'</i>				<i>c</i>			
	<i>M</i>			<i>SEM</i>	<i>M</i>			<i>SEM</i>
Short retention interval								
Short	1.62	τ	τ	0.07	-.26	τ	τ	0.04
		n				n		
Long	1.51	τ	⊥	0.06	-.26	τ	⊥	0.04
		n				***		
Strong	1.43	⊥	⊥	0.08	0.00	⊥	⊥	0.04
Long retention interval								
Short	1.08	τ	τ	0.07	-.22	τ	τ	0.04
		n				n		
Long	1.00	τ	⊥	0.08	-.20	τ	⊥	0.05
		n				***		
Strong	1.01	⊥	⊥	0.08	-.04	⊥	⊥	0.04

Note. Normal distribution assumed. *n* non-significant; \*\*\*  $p < .001$ .  $N = 96$ .

Table A.31. Sensitivity ( $d'$ ) across lure types.

List type	$d'$ (SP lures)		$d'$ (unr. lures)	
	$M$	$SEM$	$M$	$SEM$
<b>Short retention interval</b>				
Short	1.08	$\tau \tau$	0.08	2.04 $\tau \tau$ 0.07
Long	0.97	$\tau \perp$ $n^{**}$	0.07	1.95 $\tau \perp$ $n$ 0.07
Strong	0.86	$\perp \perp$ $n$	0.07	1.96 $\perp \perp$ $n$ 0.09
<b>Long retention interval</b>				
Short	1.08	$\tau \tau$	0.07	1.92 $\tau \tau$ 0.08
Long	1.00	$\tau \perp$ $n$	0.08	1.88 $\tau \perp$ $n$ 0.07
Strong	1.01	$\perp \perp$ $n$	0.08	1.94 $\perp \perp$ $n$ 0.08

Note. unr. = unrelated.  $n$  non-significant; ;  $** p < .01$ .  $N = 96$ .

Table A.32. Sensitivity ( $A_z$ ) across encoding tasks and comparison types.

List type	Short interval		Long interval	
	$M$	$SEM$	$M$	$SEM$
<b>Studied vs. Unrelated lures</b>				
Short	.89	$\tau \tau$ *	.86	$\tau \tau$ $n$
Long	.87	$\tau \perp$ *	.86	$\tau \perp$ $n$
Strong	.85	$\perp \perp$ $n$	.85	$\perp \perp$ $n$
<b>Studied vs. Switched plurality</b>				
Short	.76	$\tau \tau$ $n^{**}$	.74	$\tau \tau$ $n$
Long	.74	$\tau \perp$ *	.74	$\tau \perp$ $n$
Strong	.70	$\perp \perp$ $n$	.73	$\perp \perp$ $n$
<b>Switched plurality vs. Unrelated</b>				
Short	.64	$\tau \tau$ $n$	.64	$\tau \tau$ $n$
Long	.66	$\tau \perp$ *	.65	$\tau \perp$ $n^{**}$
Strong	.70	$\perp \perp$ $n$	.68	$\perp \perp$ $n$

Note.  $A_z$  = area under the ROC;  $n$  non-significant;  $\dagger p < .10$ ;  $* p \leq .05$ ;  $** p < .01$ .  $N = 72$ .

**Table A.33. Bias ( $c_a$ ) across encoding tasks and comparison types.**

List type	Short interval				Long interval			
	<i>M</i>		<i>SEM</i>		<i>M</i>		<i>SEM</i>	
Studied vs. unrelated lures								
Short	0.18		† n	† ***	0.04	0.20	† n	† **
Long	0.19	† ***	⊥		0.04	0.25	† **	⊥
Strong	0.52	⊥		⊥	0.05	0.37	⊥	⊥
Studied vs. switched plurality								
Short	-.25		† n	† ***	0.04	-.19	† †	† ***
Long	-.24	† ***	⊥		0.04	-.12	† *	⊥
Strong	0.05	⊥		⊥	0.05	-.02	⊥	⊥
Switched plurality vs. unrelated								
Short	0.65		† n	† ***	0.05	0.68	† n	† ***
Long	0.62	† ***	⊥		0.05	0.69	† **	⊥
Strong	0.92	⊥		⊥	0.04	0.85	⊥	⊥

Note.  $c_a$  = response bias (from  $X_3$ );  $n$  non-significant; †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ .

## Appendix 2

In this Appendix, we analyse response time data. Both *encoding times* (time to enter a response at study) and *retrieval times* (the time to enter a response at test) are reported. Encoding times are relevant in Experiments 1 to 4 because responses to the encoding tasks are self-paced in those studies. Consequently, it is important to assess in those studies whether the amount of time spent encoding study items was equivalent across list types. Retrieval times are also reported; the conclusions derived from retrieval times, however, are not emphasised in the main text. That is because participants were instructed to favour accuracy over speed. As a result, retrieval times may not appropriately reflect the speed of retrieval.

### Experiment 1

The use of a self-paced encoding task in Experiments 1 to 4 naturally introduces a confound: the average study time for *strong* lists was shorter than the study time for *long* lists, because participants responded faster to repeated words in *strong* lists. Average encoding times were indeed shorter for *strong* lists ( $M = 222$  s,  $SEM = 2.8$ ) compared to *long* lists ( $M = 231$  s,  $SEM = 2.5$ ),  $t(67) = 3.12$ ,  $SEM = 2.8$ ,  $p < .01$ . More importantly, however, no difference was found between average encoding times for *target* items (*short*:  $M = 1.76$  s,  $SEM = 0.02$ ; *long*:  $M = 1.76$  s,  $SEM = 0.03$ ; *strong*:  $M = 1.77$  s,  $SEM = 0.03$ ;  $F < 1$ ,  $p = .96$ ), confirming that *targets* were studied for the same average amount of time in all three list types. Moreover, there was no difference in target encoding times across encoding tasks,  $F < 1$ ,  $p = .80$ , suggesting that the encoding tasks did not differ in terms of difficulty.

At test, response times did differ across encoding tasks: it took longer to make a decision about targets when the item was encoded in the *size* condition ( $M = 1.81$  s,  $SEM = 0.05$ ) than when the item was studied in the *pleasantness* condition ( $M = 1.67$  s,  $SEM = 0.05$ ),  $F(1,66) = 3.91$ ,  $MSE = 0.25$ ,  $p = .05$ . Retrieval times did not differ across list types and there was no interaction between list type and encoding task ( $F_s < 1.6$ ,  $p_s > .21$ ). The interaction between experiment phase (*study*, *test*) and task (*size*, *pleasantness*) was significant,  $F(1,132) = 4.77$ ,  $MSE = 0.15$ .

## Experiment 2

Average encoding times were shorter for *strong* lists ( $M = 225$  s,  $SEM = 3.03$ ) compared to *long* lists ( $M = 236$  s,  $SEM = 2.63$ ),  $t(113) = 4.72$ ,  $SEM = 2.3$ ,  $p < .001$ . Unexpectedly, there was also a significant difference in the average encoding times of *targets* across lists,  $F(2,226) = 3.51$ ,  $MSE = 29.62$ ,  $p = .03$ , such that targets were studied for less time in *short* lists ( $M = 1.80$  s,  $SEM = 0.02$ ) than in *long* ( $M = 1.86$  s,  $SEM = 0.02$ ) and *strong* lists ( $M = 1.86$  s,  $SEM = 0.03$ ). There was no difference at test in retrieval times across list types,  $F < 1$ ,  $p = .38$ . The interaction between experiment phase (*study* vs. *test*) and list type was significant,  $F(2,226) = 3.97$ ,  $MSE = 42.71$ ,  $p = .02$ . The latter result may be interpreted as evidence that the shorter average encoding time that accompanied *short* lists had little detrimental effect on recognition performance: despite studying each target item for about 60 ms less in *short* lists, participants took the same time on average to make a recognition decision at test for *short*, *long* and *strong* lists.

## Experiment 3

Average encoding times were shorter for *strong* lists ( $M = 225$  s,  $SEM = 2.85$ ) compared to *long* lists ( $M = 234$  s,  $SEM = 3.20$ ),  $t(118) = 4.37$ ,  $SEM = 2.2$ ,  $p < .001$ . The encoding times for *target* items was the same for *short* ( $M = 1.80$  s,  $SEM = 0.03$ ), *long* ( $M = 1.82$  s,  $SEM = 0.03$ ) and *strong* lists ( $M = 1.82$  s,  $SEM = 0.02$ ),  $F_s < 1$ ,  $p > .48$ . Retrieval times, on the other hand, differed slightly across list types,  $F(2,234) = 2.44$ ,  $MSE = 0.11$ ,  $p = .09$ , so that responses were faster for *short* lists. Retrieval times also differed slightly across encoding tasks,  $F(1,117) = 3.85$ ,  $MSE = 0.60$ ,  $p = .05$ , such that responses were shorter in the *size* condition.

## Experiment 4

Average encoding times were shorter to *strong* lists ( $M = 351$  s,  $SEM = 4.57$ ) than to *long* lists ( $M = 391$  s,  $SEM = 4.35$ ),  $t(99) = 10.26$ ,  $SEM = 3.8$ ,  $p < .001$ ; there was no effect of retention interval and no interaction with list type. Surprisingly, participants took slightly more time, on average to encode *targets* in *long* lists ( $M = 1.81$  s,  $SEM = 0.02$ ) than in *short* ( $M = 1.76$  s,  $SEM = 0.03$ ) and *strong* lists ( $M = 1.78$  s,  $SEM = 0.03$ ),  $F(2,196) = 2.66$ ,  $MSE = 0.03$ ,  $p = .08$ . There was no effect of



retention interval and no interaction,  $F_s < 1$ ,  $ps > .32$ . There was also no main effect of list type and retention interval on retrieval times,  $F_s < 1.2$ ,  $ps > .30$ . There was, however, an interaction between the two variables,  $F(2,196) = 4.0$ ,  $MSE = 0.11$ , suggesting that the speed of recognition did not vary across lists when retention interval was long but that participants were faster at responding to *short* lists compared to *strong* lists when retention interval was short.

## Experiment 5a

For *A* items, there was no difference in the average encoding times of *targets* across list types and retention intervals (all  $F_s < 1.6$ , all  $ps > .22$ ;  $M_{enc} = 0.78$  s,  $SEM = 0.01$ ). The result shows that the use of a fixed encoding time of 1.15 s was not only effective in controlling overall study-test lag across list types but it was also effective in controlling average *target* encoding times across conditions. There was a marginal difference in the average retrieval times for *A targets*,  $F(2,94) = 2.63$ ,  $MSE = 0.10$ ,  $p = .08$ , such that retrieval of *targets* was faster in *short* lists ( $M_{ret} = 2.44$  s,  $SEM = 0.01$ ) than in *long* lists ( $M_{ret} = 2.52$  s,  $SEM = 0.01$ ) and *strong* lists ( $M_{ret} = 2.54$  s,  $SEM = 0.01$ ). There was also a nearly significant interaction between list type and retention interval in the retrieval times of *A SP lures*,  $F(2,94) = 2.82$ ,  $MSE = 0.08$ ,  $p = .07$ , suggesting that participants took slightly longer to respond to *SP lures* in *long* lists but only when retention interval was short.

For *B* items, there was no main effect of average encoding times of *targets* across list type and retention intervals (all  $F_s < 1$ , all  $ps > .40$ ;  $M_{enc} = 0.78$  s,  $SEM = 0.01$ ).<sup>1</sup> The result indicates that *A* and *B* items, which were presented early on the study list, were indeed indistinguishable to participants ( $F < 1$ ,  $p = .91$ ). There was a main effect in the average retrieval times across list types,  $F(2,94) = 10.32$ ,  $MSE = 0.11$ ,  $p < .001$ , showing that participants were faster at entering responses to *B targets* in *short* ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ) and *strong* lists ( $M_{ret} = 2.3$  s,  $SEM = 0.01$ ) than in *long* lists ( $M_{ret} = 2.5$  s,  $SEM = 0.01$ ). The same pattern was observed on the retrieval of *B SP lures*, though the effect was attenuated ( $p = .07$ ).

---

<sup>1</sup> When *repeated B* items are included in the analysis, then there is a difference across lists,  $F(2,94) = 16.47$ ,  $MSE = 0.03$ ,  $p < .001$ , whereby *B* items are encoded faster in *strong* lists ( $M_{enc} = 0.74$  s,  $SEM = 0.01$ ) than in *short* ( $M_{enc} = 0.77$  s,  $SEM = 0.01$ ) and *long* ( $M_{enc} = 0.78$  s,  $SEM = 0.01$ ) lists.

More interesting, however, are the comparisons between retrieval times for *targets* and *SP lures* across retention intervals and list types. For *A* items, there was a strong effect of word type,  $F(1,47) = 21.13$ ,  $MSE = 0.07$ ,  $p < .001$ , showing that participants took approximately 100 ms longer to respond to *A SP lures* ( $M_{ret} = 2.6$  s,  $SEM = 0.01$ ) than to *A targets* ( $M_{ret} = 2.5$  s,  $SEM = 0.01$ ). The result is consistent with participants adopting a recall-to-reject strategy at test (i.e., presumably participants first try to recall whether the test items was studied). Moreover, there was a marginal effect of list type,  $F(2,94) = 2.89$ ,  $MSE = 0.17$ ,  $p = .06$ , indicating that responses in *short* lists were faster than responses in *long* and *strong* lists.

Similarly for *B* items, there was an effect of word type,  $F(1,47) = 12.93$ ,  $MSE = 0.11$ ,  $p = .001$ , showing that participants took 100 ms longer to respond to *B SP lures* ( $M_{ret} = 2.5$  s,  $SEM = 0.01$ ) than to *B targets* ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ). Moreover, there was an effect of list type,  $F(2,94) = 7.44$ ,  $MSE = 0.17$ ,  $p = .001$ , showing that responses in *short* and *strong* lists were faster than responses in *long* lists (note that *B* items were repeated in *strong* lists).

Overall, the results from response times suggest that retention intervals somewhat affected response times. In addition, the trends across list types indicate that length and strength manipulations were strong enough to delay participants' responses, despite not being strong enough to affect sensitivity.

## Experiment 5b

For *A* items, there was no difference in the average encoding times of *targets* across list types and retention intervals (all  $F$ s  $< 1$ , all  $p$ s  $> .58$ ;  $M_{enc} = 0.80$  s,  $SEM = 0.02$ ). There was also no difference in the average retrieval times for *A targets*, no effect of retention interval and no interaction ( $F$ s  $< 1.7$ ,  $p$   $> .20$ ). There was no difference in retrieval times for *A SP lures* across lists and no effect of retention interval ( $F$ s  $< 1$ ,  $p$ s  $> .83$ ). There was, however, a nearly significant interaction between list type and retention interval,  $F(2,46) = 3.18$ ,  $MSE = 0.07$ ,  $p = .05$ , hinting that participants were slower to respond to *SP lures* in *strong* lists in the

short retention interval condition but faster in the long interval condition compared to *short* and *strong* lists.

We also compared retrieval times across word types (*targets* vs. *SP lures*). There was a strong effect of word type,  $F(1,23) = 22.65$ ,  $MSE = 0.07$ ,  $p < .001$ , showing that participants took about 150 ms longer to respond to *A SP lures* ( $M_{ret} = 2.6$  s,  $SEM = 0.01$ ) than to *A targets* ( $M_{ret} = 2.7$  s,  $SEM = 0.01$ ). The result is consistent with the use of a recall-to-reject strategy at test. Moreover, there was an interaction between word type and retention interval,  $F(1,23) = 7.34$ ,  $MSE = 0.04$ , indicating that responses to *targets* decreased from short to long retention intervals whereas responses to *SP lures* increased.

For *B* items, there was no effect of average encoding times of *targets* across list type and retention intervals (all  $F$ s  $< 1$ , all  $p$ s  $> .47$ ;  $M_{enc} = 0.80$  s,  $SEM = 0.02$ ).<sup>2</sup> The result indicates that *A* and *B* items were indistinguishable to participants ( $F < 1$ ,  $p = .67$ ). The average retrieval times for *B targets* were shorter in *strong* lists ( $M_{ret} = 2.3$  s,  $SEM = 0.01$ ) than in *short* ( $M_{ret} = 2.5$  s,  $SEM = 0.01$ ) and *long* ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ) lists,  $F(2,46) = 2.78$ ,  $MSE = 0.09$ ,  $p = .07$ . The same pattern was observed on the retrieval of *B SP lures*,  $F(2,46) = 3.78$ ,  $MSE = 0.07$ ,  $p = .03$ .

Analyses of retrieval times across word types (*targets* vs. *SP lures*) yielded an effect of word type,  $F(1,23) = 17.44$ ,  $MSE = 0.11$ ,  $p < .001$ , showing that participants took 170 ms longer to respond to *B SP lures* ( $M_{ret} = 2.6$  s,  $SEM = 0.01$ ) than to *B targets* ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ). Moreover, there was an effect of list type,  $F(2,46) = 3.93$ ,  $MSE = 0.13$ , showing that responses were faster in *strong* lists ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ) than in *short* ( $M_{ret} = 2.5$  s,  $SEM = 0.01$ ) and *long* ( $M_{enc} = 2.5$  s,  $SEM = 0.01$ ) lists. The result indicates that participants were able to reach a recognition decision quicker to both *targets* and *SP lures* when the test items were strong.

---

<sup>2</sup> When *repeated B* items are included in the analysis, then there is a difference across lists,  $F(2,46) = 20.41$ ,  $MSE = 0.02$ ,  $p < .001$ , whereby *B* items are encoded faster in *strong* lists ( $M_{enc} = 0.75$  s,  $SEM = 0.01$ ) than in *short* ( $M_{enc} = 0.80$  s,  $SEM = 0.02$ ) and *long* ( $M_{enc} = 0.80$  s,  $SEM = 0.02$ ) lists.

## Experiment 6

For *A* items, there was no difference in the average encoding times of *targets* across list types and retention intervals (all  $F_s < 1$ , all  $p_s > .57$ ;  $M_{enc} = 0.76$  s,  $SEM = 0.01$ ). There was also no difference in the average retrieval times for *A targets*, no effect of retention interval and no interaction ( $F_s < 1.1$ ,  $p > .34$ ). There was no difference in retrieval times for *A SP lures* across lists, no effect of retention interval and no interaction ( $F_s \leq 1$ ,  $p_s > .37$ ). Thus, length, strength and retention interval manipulations did not affect encoding times for *A targets* and did not affect response times for both *A targets* and *A SP lures*.

Retrieval times were also compared across word types (*targets* vs. *SP lures*). If participants take reliably longer to respond to *SP lures* compared to *targets*, this may suggest that they are using a recall-to-reject strategy: participants would fail to recall a test item when it is an *SP lure* more often than when it is a *target* and, consequently, would take longer to respond to *SP lures*. Indeed, there was a strong effect of word type,  $F(1,47) = 13.95$ ,  $MSE = 0.08$ ,  $p = .001$ , showing that participants took about 100 ms longer to respond to *A SP lures* ( $M_{ret} = 2.5$  s,  $SEM = 0.01$ ) than to *A targets* ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ).

For *B* items, there was no effect of average encoding times of *targets* across list type and retention intervals ( $F_s < 1.47$ ,  $p_s > .24$ ;  $M_{enc} = 0.76$  s,  $SEM = 0.01$ ).<sup>3</sup> The result indicates that *A* and *B* items were indistinguishable to participants: there was no difference in response times between *A* and *B* targets ( $F < 1$ ,  $p = .61$ ). The average retrieval times for *targets* were shorter in *strong* lists ( $M_{ret} = 2.1$  s,  $SEM = 0.01$ ) than in *short* ( $M_{ret} = 2.3$  s,  $SEM = 0.01$ ) and *long* ( $M_{ret} = 2.3$  s,  $SEM = 0.01$ ) lists,  $F(2,94) = 16.12$ ,  $MSE = 0.12$ ,  $p < .001$ . This is consistent with the fact that *targets* in *strong* lists were more strongly encoded due to repetition. There was no difference in response times for *B SP lures* across list types ( $F < 1$ ,  $p = .66$ ).

Analyses of retrieval times across word types (*targets* vs. *SP lures*) yielded an effect of word type,  $F(1,47) = 59.09$ ,  $MSE = 0.01$ ,  $p < .001$ , showing that

<sup>3</sup> When *repeated B* items are included in the analysis, then there is a difference across lists,  $F(2,94) = 66.78$ ,  $MSE = 0.01$ ,  $p < .001$ , whereby *B* items are encoded faster in *strong* lists ( $M_{enc} = 0.68$  s,  $SEM = 0.01$ ) than in *short* ( $M_{enc} = 0.76$  s,  $SEM = 0.01$ ) and *long* ( $M_{enc} = 0.77$  s,  $SEM = 0.01$ ) lists.

participants took about 60 ms longer to respond to *B SP lures* ( $M_{ret} = 2.34$  s,  $SEM = 0.01$ ) than to *B targets* ( $M_{ret} = 2.28$  s,  $SEM = 0.01$ ). Moreover, there was an effect of list type,  $F(2,94) = 6.15$ ,  $MSE = 0.21$ ,  $p = .003$ , showing that responses were faster in *strong* lists ( $M_{ret} = 2.2$  s,  $SEM = 0.01$ ) than in *short* ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ) and *long* ( $M_{enc} = 2.3$  s,  $SEM = 0.01$ ) lists. Finally, there was an interaction between word type and list type,  $F(2,94) = 13.82$ ,  $MSE = 0.05$ ,  $p < .001$ . The interaction indicates that participants were able to reach a recognition decision quicker to *targets* (but not to *SP lures*) when the item being tested was repeated at study.

## Experiment 7

Encoding times did not differ across list types and across retention intervals,  $F_s < 1$ ,  $p_s > .39$ ,  $M_{enc} = 0.76$  s,  $SEM = 0.01$ . When repeated items in *strong* lists were included in the analysis, a difference in encoding times across lists emerged,  $F(2,142) = 77.67$ ,  $MSE = 0.03$ ,  $p < .001$ , whereby study items were, on average, encoded faster in *strong* lists ( $M_{enc} = 0.69$  s,  $SEM = 0.01$ ) than in *short* ( $M_{enc} = 0.76$  s,  $SEM = 0.01$ ) and *long* ( $M_{enc} = 0.76$  s,  $SEM = 0.01$ ) lists.

Retrieval times were analysed separately for *targets*, *SP lures* and *unrelated lures*. For *targets*, there was no effect of retention interval and no interaction with list type,  $F_s < 2.0$ ,  $p_s > .14$ . There was, however, a marginal main effect of list type,  $F(2,142) = 2.39$ ,  $MSE = 0.19$ ,  $p = .09$ , hinting that participants took slightly longer to retrieve *targets* in *long* lists ( $M_{ret} = 2.55$  s,  $SEM = 0.01$ ) than in *short* ( $M_{ret} = 2.45$  s,  $SEM = 0.01$ ) and *strong* lists ( $M_{ret} = 2.45$  s,  $SEM = 0.01$ ). A similar pattern was found for *SP lures*: no main effect of retention interval and no interaction with list type,  $F_s < 1$ ,  $p_s > .38$  but a marginal main effect of list type,  $F(2,142) = 2.80$ ,  $MSE = 0.18$ ,  $p = .06$ ; participants took longer to retrieve *targets* in *long* lists ( $M_{ret} = 2.70$  s,  $SEM = 0.01$ ) than in *short* ( $M_{ret} = 2.62$  s,  $SEM = 0.01$ ) and *strong* lists ( $M_{ret} = 2.59$  s,  $SEM = 0.01$ ). Similarly, for *unrelated lures*, participants took longer to respond in *long* lists ( $M_{ret} = 2.30$  s,  $SEM = 0.01$ ) than in *short* ( $M_{ret} = 2.22$  s,  $SEM = 0.01$ ) and *strong* lists ( $M_{ret} = 2.07$  s,  $SEM = 0.01$ ) and took longer to respond in *short* lists than in *strong* lists,  $F(2,142) = 19.10$ ,  $MSE = 0.10$ ,  $p < .001$ .

Retrieval times were also compared across word types (*targets*, *SP lures* and *unrelated lures*). There was a strong effect of word type,  $F(2,142) = 195.69$ ,  $MSE = 0.11$ ,  $p < .001$ , showing that responses to *SP lures* ( $M_{ret} = 2.6$  s,  $SEM = 0.01$ ) took about 100 ms longer than responses to *targets* ( $M_{ret} = 2.5$  s,  $SEM = 0.01$ ) which in turn took about 300 ms longer than responses to *unrelated lures* ( $M_{ret} = 2.2$  s,  $SEM = 0.01$ ). There was also a main effect of list type,  $F(2,142) = 7.22$ ,  $MSE = 0.32$ ,  $p = .001$ , showing that it took longer to respond to test items in *long* lists ( $M_{ret} = 2.5$  s,  $SEM = 0.01$ ) than to respond to test items in *short* ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ) and *strong* lists ( $M_{ret} = 2.4$  s,  $SEM = 0.01$ ). Finally, there was an interaction between word type and list type,  $F(4,284) = 3.85$ ,  $MSE = 0.07$ ,  $p = .005$ , such that response times to *targets* and *SP lures* were similar for *short* and *strong* lists but response times for *unrelated lures* were shorter in *strong* lists.

## Appendix 3

Here we present the proportion hits and false alarms that were corrected to avoid infinite  $d'$  values. If  $H = 1$ , then  $H_{\text{corr}} = 1 - 1/2n$ , where  $n$  is the number of test trials per word type (i.e., *target*, *SP lure*, *unrelated lure*;  $n = 15$  or  $30$ ). If  $FA = 0$ , then  $FA_{\text{corr}} = 1/2n$  (Macmillan and Creelman, 2005). The number of participants whose data was corrected is given in brackets in the tables below.

**Table A.34. Proportion of corrected hits and false alarms (number of participants in brackets).**

<b>Experiment 1</b>						
<b>List type</b>	<b>Size</b>			<b>Pleasantness</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.17 (6)	-	.11 (4)	.44 (16)	-	.11 (4)
<b>Long</b>	.06 (2)	-	.06 (2)	.31 (11)	-	.17 (6)
<b>Strong</b>	.00 (0)	-	.19 (7)	.17 (6)	-	.22 (8)

<b>Experiment 2</b>			
<b>List type</b>	<b>Size</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.06 (8)	.00 (0)	.13 (17)
<b>Long</b>	.07 (9)	.01 (1)	.14 (18)
<b>Strong</b>	.06 (7)	.02 (2)	.28 (35)

*HR* = hits; *SP-FAR* = false-alarms (switched plurality); *FAR* = false-alarms (unrelated lures).

**Table A.34. Proportion of corrected hits and false alarms (continued).**

<b>Experiment 3</b>						
<b>List type</b>	<b>Size</b>			<b>Pleasantness</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.06 (4)	.03 (2)	.21 (14)	.14 (9)	.00 (0)	.11 (7)
<b>Long</b>	.08 (5)	.02 (1)	.18 (12)	.12 (8)	.03 (2)	.12 (8)
<b>Strong</b>	.02 (1)	.02 (1)	.30 (20)	.17 (11)	.00 (0)	.21 (14)

<b>Experiment 4</b>						
<b>List type</b>	<b>Short interval</b>			<b>Long interval</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.11 (6)	.00 (0)	.09 (5)	.06 (3)	.02 (1)	.04 (2)
<b>Long</b>	.02 (1)	.00 (0)	.04 (2)	.06 (3)	.00 (0)	.04 (2)
<b>Strong</b>	.02 (1)	.00 (0)	.15 (8)	.06 (3)	.00 (0)	.09 (5)

*HR* = hits; *SP-FAR* = false-alarms (switched plurality); *FAR* = false-alarms (unrelated lures).



Table A.34. Proportion of corrected hits and false alarms (continued).

<b>Experiment 5a</b>						
<b>A items</b>						
<b>List type</b>	<b>Short interval</b>			<b>Long interval</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.04 (2)	.04 (2)	-	.04 (2)	.02 (1)	-
<b>Long</b>	.00 (0)	.04 (2)	-	.04 (2)	.00 (0)	-
<b>Strong</b>	.02 (1)	.02 (1)	-	.00 (0)	.00 (0)	-
<b>B items</b>						
<b>List type</b>	<b>Short interval</b>			<b>Long interval</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.04 (2)	.04 (2)	-	.00 (0)	.04 (2)	-
<b>Long</b>	.02 (1)	.04 (2)	-	.04 (2)	.06 (3)	-
<b>Strong</b>	.23 (11)	.08 (4)	-	.13 (6)	.08 (4)	-
<b>Experiment 5b</b>						
<b>A items</b>						
<b>List type</b>	<b>Short interval</b>			<b>Long interval</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.08 (2)	.00 (0)	-	.00 (0)	.04 (1)	-
<b>Long</b>	.00 (0)	.00 (0)	-	.00 (0)	.00 (0)	-
<b>Strong</b>	.08 (2)	.04 (1)	-	.00 (0)	.00 (0)	-
<b>B items</b>						
<b>List type</b>	<b>Short interval</b>			<b>Long interval</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.17 (4)	.06 (2)	-	.04 (1)	.00 (0)	-
<b>Long</b>	.00 (0)	.03 (1)	-	.00 (0)	.04 (1)	-
<b>Strong</b>	.29 (7)	.17 (4)	-	.17 (4)	.08 (2)	-

*HR* = hits; *SP-FAR* = false-alarms (switched plurality); *FAR* = false-alarms (unrelated lures).

Table A.34. Proportion of corrected hits and false alarms (continued).

<b>Experiment 6</b>						
<b>A items</b>						
<b>List type</b>	<b>Short interval</b>			<b>Long interval</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.03 (1)	.08 (3)	-	.06 (2)	.08 (3)	-
<b>Long</b>	.08 (3)	.08 (3)	-	.03 (1)	.03 (1)	-
<b>Strong</b>	.00 (0)	.11 (4)	-	.03 (1)	.08 (3)	-
<b>B items</b>						
<b>List type</b>	<b>Short interval</b>			<b>Long interval</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.03 (1)	.06 (2)	-	.03 (1)	.03 (1)	-
<b>Long</b>	.03 (1)	.03 (1)	-	.03 (1)	.02 (1)	-
<b>Strong</b>	.11 (4)	.17 (6)	-	.11 (4)	.17 (6)	-
<b>Experiment 7</b>						
<b>List type</b>	<b>Short interval</b>			<b>Long interval</b>		
	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>	<i>HR</i>	<i>SP-FAR</i>	<i>FAR</i>
<b>Short</b>	.05 (5)	.01 (1)	.07 (7)	.04 (4)	.01 (1)	.08 (8)
<b>Long</b>	.04 (4)	.01 (1)	.09 (9)	.05 (5)	.02 (2)	.06 (6)
<b>Strong</b>	.00 (0)	.02 (2)	.24 (23)	.03 (3)	.03 (3)	.16 (15)

*HR* = hits; *SP-FAR* = false-alarms (switched plurality); *FAR* = false-alarms (unrelated lures).